

CANADIAN TRANSPORT COMMISSION
Research Seminar Series
Volume 8
Spring 1982

Errata

On Page 7, the fourth line of the third paragraph, the word "log" should read "logit".

Pages 15 and 17 should be reversed.

On the top of page 46, the author's name should read "Christensen, L.R." rather than "Caves, L.R."

RESEARCH SEMINAR SERIES

**Richard H. Fosbrooke
Series Co-ordinator**

**Hope Dempsey
Editor, English Version**

**Louise Lambert
Editor, French Version**

**CANADIAN TRANSPORT COMMISSION
Research Branch
Spring 1982
Volume 8**

E
On
th
sh
Pa
On
nar
rat

Version française disponible sous le titre
"Série de séminaires de recherche",
volume 8, printemps 1982

First Printing, May 1983

Canadian Transport Commission
15 Eddy St., 15th floor, Ottawa-Hull K1A 0N9

©Minister of Supply and Services Canada 1983
Catalogue Number TT 71-21/1982-1E
ISBN 0-662-92165-8
ISSN 0229-7574

Printed in Canada



FOREWORD

This is the text of two seminars that were presented to the Commissioners and staff of the Canadian Transport Commission during the Spring of 1982. The Commission makes these papers available as a further contribution to the continuing discussion of the development of transportation policy and regulation in Canada.

The seminar presented by W.G. Waters II, Professor at the University of British Columbia, addresses some of the recent developments in the economics of transport regulation. These advances in economic theory with respect to the role of competition in transportation markets have indicated a number of complexities, alternative interpretations and emphasis which together suggest a changing role for government interference in market forces.

The address by M. Florian, Professor at the Centre de recherche sur les transports (Université de Montréal), describes the role of network models in transportation planning and their interaction/graphic implementation. The presentation concentrates on the elements of network models used in transportation planning, reviews their structural properties and the most commonly used solution methods, and outlines their current and potential applications.

It must be taken for granted that the opinions expressed in these papers are those of the respective authors and do not necessarily reflect the views of the Canadian Transport Commission.

Yves Dubé
Vice-President (Research)

CHURCH

CONTENTS

RECENT DEVELOPMENTS IN THE ECONOMICS OF TRANSPORT REGULATION

W.G. Waters II

1

NETWORK MODELS IN TRANSPORTATION PLANNING AND THEIR INTERACTIVE/GRAPHIC IMPLEMENTATION

Michael Florian

49



**RECENT DEVELOPMENTS IN THE ECONOMICS
OF TRANSPORT REGULATION**

by

**W.G. Waters II
Faculty of Commerce and Business Administration
University of British Columbia**



RECENT DEVELOPMENTS IN THE ECONOMICS OF TRANSPORT REGULATION*

Much of Transport Economics has been a relatively timeless subject. Of course, nothing is ever timeless literally, but so many issues and controversies have a familiar ring: controversies over estimating costs when joint and common costs and indivisibilities abound; trying to determine if a particular set of regulations will improve or harm economic efficiency; trying to estimate efficient prices for publicly-provided facilities, etc. Underlying these and other controversies has been the solid base of microeconomic theory and applied welfare economics, albeit with modification or special emphasis to reflect salient characteristics of the provision of transportation services, e.g., multiple products and joint and common costs, indivisibilities, non-storable outputs, possible increasing returns, externalities, etc. Refinements to the theory and advances in empirical methods are forthcoming regularly. But these refinements and improved techniques of analysis have been just that, they have not fundamentally altered our understanding of and thinking about the workings of transport markets. In the absence of externalities, public goods or some other "market failure", competition is generally regarded as the desired method for achieving efficient allocation of resources. Where natural monopoly is present, government regulation can be called upon to intervene and simulate the results of competition (price equals marginal cost, etc.).

In the last few years some theoretical arguments have appeared which question our traditional confidence in economic performance of competition in transportation markets. Paradoxically, there are some other arguments which suggest that regulation may not be necessary even in cases of monopoly. Like any major revision to a body of theory, new developments do

* Paper accompanied a seminar given to the Canadian Transport Commission on April 30, 1982. The author wishes to acknowledge the helpful comments by and conversations with his colleagues: Kenneth J. Button, T.D. Heaver and Steven A. Morrison. Thanks to Ken Freeman for his helpful research assistance. The author, of course, bears full responsibility for the analysis and opinions expressed herein.

not show previous theories to be wrong per se but draw attention to complexities and alternative interpretations and emphasis which, nonetheless, lead to different conclusions than previously. It is really too early to tell if these developments are as revolutionary as some might suggest. Some of these arguments point to much greater interference in market forces by government regulations. It is ironic that such conclusions would come just at a time when decades of criticism of regulation by economists have borne fruit with the deregulation movement in North America.

The object of this paper is to summarize some recent theoretical analyses which have appeared in major economics journals and which are directly relevant to transport regulation. Theory deals with precise formulations usually in mathematical form. This degree of rigor is necessary and appropriate for deducing precisely the implications of theoretical postulates. But there is a wide gulf between the precision and elegance of theoretical works compared to the more complex "real world" of transport decision making and regulation. This paper attempts to provide a relatively concise and simple explanation of some of these recent developments on the frontiers of economic research, and their potential implications for transport regulation.

Before turning to the topics of this paper, there are a few additional introductory remarks which are in order. There are a good many "recent developments" which could be discussed; I can deal with only a few. It may be helpful if I first clarify my understanding of the major interests and thrust of the economics of transport regulation, i.e., just which parts of economic theory are most relevant to the topic. Second, there are a number of interesting and important "recent developments" which cannot be reviewed here. I wish to acknowledge some of these topics prior to getting down to the ones being emphasized here.

A. What is Transport Economics?

I could begin with the truism that transportation economics is whatever transport economists say and do. This would encompass a gamut of economic theory and methods from the most practical efforts at forecasting and

costing to rigorous analysis and estimation of systems of equations. My interest today is in the more conceptual aspects of transport economics, but I promise not to dazzle you with systems of equations.

It is interesting that what constitutes the major issues in transport economics differs somewhat from one country to another. In the U.K., the main thrust appears to be on identifying pricing and investment rules for the public sector, and topics derivative to this. The emphasis in English journals and reports is on the economic evaluation of public transport projects (e.g., social benefit cost studies) and efficient pricing in the presence of externalities and substitution effects among modes. The major topic is urban transportation but high-density inter-urban travel receives attention as well. Derivative issues include the estimation of travel demand including cross elasticities between modes and the elasticities with respect to travel time and related quality variables. Estimating the value of travel time has received special emphasis. It is important both for estimating modal splits and also because time savings constitute the major economic benefit of the majority of public transport projects. Indeed, one prominent English author maintains that establishing the value of travel time is one of the most important issues in transport economics (Sharp, 1973).

In North America, particularly in the U.S., the emphasis of transportation economics has always been much more on regulatory matters. More generally, I think the main themes have focussed on the problems of achieving efficiency in multiproduct enterprises whether by market competition or using regulation as a substitute for free markets. Regulation refers primarily to price and/or entry controls. Transport economics texts review the problems of costing and pricing in the presence of joint and common costs, the need for regulation in the presence of natural monopoly or other imperfectly-competitive market structures, and the guidelines for efficient pricing under these circumstances. North American texts normally include considerable attention to existing regulatory

procedures and their economic shortcomings. Evaluation of public investments traditionally receives much less attention and is addressed primarily on the issue of user charges for publicly-provided facilities.

In Canada, the traditional issues of emphasis in transport economics have generally been the same as the U.S., but with greater attention to the promotional role of government in providing transport facilities. This is not surprising given the greater presence of government in the provision of transport facilities in Canada. This was necessitated by our sparse population and the importance of transportation for political union.

Australian interests in transport economics share the English emphasis on evaluation of public pricing and investment rules and a strong emphasis on urban transportation. But Australia also shares Canada's concern over the roles of transportation in promoting economic development (although Australia places greater emphasis on ocean transportation in contrast to the traditional Canadian emphasis on rail transportation).

I am less familiar with transportation economics and policy issues on the European continent, but their emphasis appear to parallel England's stress on pricing and investment rules for the public sector. The European nations have also been concerned with the implications of differing regulatory environments and other government involvement in transportation and the inevitable problems arising from closer economic integration. Also the European nations may have a stronger tradition of political economy, i.e., emphasizing political and social forces and their influence on the economy.

My focus on "recent developments" will emphasize the orientation of transport economics which I think has been traditional in North America. These are issues which grow out of interest in the relationship between pricing and economic efficiency in multiproduct enterprises under different market structures. The relative merits of free competition versus regulation to achieve efficiency are of central concern.

B. Recent Developments in Transport Economics Not Being Reviewed

There are a large number of potential topics which could be discussed under the title heading; far too many to be covered in a single seminar. At a minimum, it is necessary to draw at least brief attention to some of the topics which will not be reviewed here.

First of all one could mention the significant advances in empirical methods which have developed over the last decade or two. I am thinking particularly of two developments. First is the formulation and estimation of multiple product cost functions capable of reflecting complex cost interrelationships among various outputs, i.e., the so-called flexible forms. These more sophisticated econometric techniques are already in widespread use, the estimate of total factor productivity for Canadian railways and airlines are just two examples which are "close to home." [1]

The second major category of empirical advances concerns demand estimation. In particular, I am thinking of the considerable work estimating mode choice primarily in passenger travel. The application of multinomial log models and similar techniques have made important advances in identifying the empirical significance of time savings and other quality attributes which are important for explaining mode choice. These studies are especially important to avoid inappropriate investments in urban transportation systems.

There have also been advances in techniques for evaluating government investments in transport facilities. I am thinking of social benefit cost analysis and related planning-evaluation techniques. No one would pretend that this has become a science, but the technique is coming into wider use and we all hope such techniques will make project evaluation, comparisons, and budget allocations easier, or at least more clear cut. [2] These techniques have not undergone any radical reformulation in recent years. The major innovations or extensions that come to mind are the increasing attention to the procedures for evaluating investments while taking into

account price-cost divergences (i.e., second best considerations) and cross effects among different demands. Here in Canada the analysis by Borins (1978) comes to mind.

Optimal pricing and investment for congested facilities is another important topic in transportation economics. The concept has been known for some time. Users of congested facilities consider their own time and inconvenience of congestion but fail to consider the equivalent costs they impose on others. The primary research on this topic has been concerned with urban transportation [3] but the topic has also been explored for airports (e.g., Carlin and Park, 1970) and ports (Vanag, 1977). Singapore has adopted a form of road pricing and there is serious interest in Kuala Lumpur and London. The major advances in this topic have been to incorporate congestion and non-optimal prices in different modes when analysing policies. On a related matter is formulation of rules for optimal investments in congested facilities in the absence of congestion prices. [4]

Considerable work has been done on the subject of "second best" prices. The basic textbook rule of $P = MC$ or "first best" pricing assumes the prices equal marginal costs elsewhere in the economy. But, as is well known, prices diverge from costs for a variety of reasons including taxes, externalities, monopoly markets, etc. Under these circumstances setting prices equal to marginal costs usually is not the optimal policy. [5] Unfortunately, no golden rules of universal validity have emerged for setting second best prices. The appropriate policy depends on the price cost divergences and various elasticities of supply and demand particular to each situation. But some guidelines have emerged and I will make reference to some in this paper. A particular type of second best pricing has received considerable attention of late. This concerns the efficiency rules for pricing to satisfy a revenue constraint such as avoiding a deficit in a decreasing cost industry. This topic will be explored below.

My main interests today are on "recent developments" in the economics of transport regulation. But this is still too broad a topic. The economics of regulation has had considerable interest in recent years and for good reason. Not only has much happened in terms of regulatory reform and

revision, but there has been a number of quite interesting developments on the underlying economic theory relevant to this topic. Let me draw attention to some recent developments in the economics of regulation relevant to transport which are not being reviewed today.

First I should remind you of extensive literature which has explored the logical implications of rate base - rate of return regulation, usually referred to as the "AJ Model" in deference to Averch and Johnson (1962). The possibility that rate of return regulation could lead to inefficient use of capital was long suspected, i.e., the AJ Model results were no great surprise, but that simple model and its progeny gave a better understanding of the sources and kinds of capital intensive biases under rate of return regulation. [6]

There has been extensive literature developed analysing the economic costs of inefficient regulation. By and large, the ideas and methods of analysis were not new, although refinements are always forthcoming. The significant development was the proliferation of estimates of the costs of regulation. [7] The findings of these numerous studies influenced the deregulation movement of the last few years.

Another significant development has been the attention focussed on alternate theories of regulation. I am thinking of the so-called "capture" theory and more sophisticated positive theories of the regulatory process (e.g., Posner, 1974, Stigler, 1971). Along the same lines, we can probably look forward to further studies of regulatory change as researchers review and try to explain what happens as different regulatory regimes (including deregulation of various degrees) take place in North America. That is, a number of regulatory experiments are taking place and we can expect to learn much from studying what happens as regulations are changed and firms and markets adapt. The theory of "contestable markets" has emerged from observations and analysis of when competition can substitute for regulation. It is safe to say that we are going to see circumstances where deregulation will not be the panacea that some of us thought it might be. And this brings me to some of the topics I wish to review.

The first topic, reviewed only briefly, is concerned with economic efficiency and competitive market equilibrium when there is quality competition. Certain types of regulation may be superior to free competition when differentiated products are involved. The second major theme for today requires more extensive discussion. It focuses on various interpretations and implications of decreasing costs. Decreasing costs or "natural monopoly" has long been one of the main rationales for government control rather than relying on a free market. The recent economics literature has drawn attention to different sources of decreasing costs than those usually discussed, as well as the various complexities which surround decreasing costs in multiproduct firms. There are some novel conclusions about the appropriate roles of government versus the market mechanism in these circumstances.

C. Market Efficiency Under Quality Competition

This topic is reviewed only briefly here. It is concerned with economic efficiency of market competition when there is quality differentiation among the services being supplied. The standard textbook model of perfect competition presumes homogeneous services are supplied. But the nature of transportation involves differentiated services. Transport supply takes place at discrete points in time and in indivisible blocks of capacity. In contrast, the demand for transport tends to arise continuously over time intervals and the quantity demanded rarely matches exactly with the amount of capacity supplied per trip. Departure/arrival times along with numerous other quality variables distinguish one mode and one carrier and one service from another. There has been much criticism of the economic performance of regulation in recent years. But to show that economic performance could be improved by deregulation is not the same as proving that total deregulation is superior to some regulation.

Like perfect competition, the standard textbook model of monopolistic competition predicts zero economic profits in equilibrium. However, because their products are differentiated, individual firms do not face perfectly elastic demand curves. A zero profit equilibrium with a downward sloping demand means the firm operates on the declining portion of its average cost

curve. That is, some excess capacity is present in equilibrium. This result is explained in any basic microeconomics text. The point is that free competition does not guarantee an optimal allocation of resources when products are differentiated. There is at least the possibility that some ideal public intervention could improve the situation.

More interesting are the complications which arise via price-quality tradeoffs. A range of price-quality combinations are possible with the same (e.g., zero) level of profits. Will competition result in the socially-optimal mix? Analyses of the (formerly) regulated U.S. Airline industry are instructive. The traditional CAB regulation controlled prices and entry. Prices were set in hopes of sustaining a financially viable industry. But airlines would compete via quality, primarily flight frequency (i.e., fleet size). This could be explained simply by the competitive nature of the industry but also because of the capital intensive bias implicit in rate of return regulation (e.g., the AJ model mentioned earlier). Whatever the explanation, the result was higher prices, greater flight frequency, and lower load factors than were socially optimal (e.g., Douglas and Miller, 1974; De Vaney, 1975). Their analysis along with that of Jordan (1970) and many others concludes that deregulation would be superior to the CAB policy. But these analyses were comparing competitive outcomes with a particular type of regulation. For the most part, these authors did not address whether or not free competition is superior to some form of regulation.

Panzar (1979) does address these issues. He postulates a simple model of the airline market. Airline costs are assumed constant. Demand is a function of price and quality variables (frequency and load factor). The full price of travel must include the marginal value of time delays passengers impose on one another (this may be zero for dense markets). He assumes that departure times are uniformly spread over the day. [8] Both the total market and that for each airline is a function of flight frequency. He finds that competitive entry would result in excess frequency, i.e., competitive entry yields higher prices, greater frequency (and lower load factors) than socially optimal. Ideal regulation would be

to set the price at the social optimum and allow free entry. Zero profits in the competitive equilibrium prevent the firms from investing in additional capacity hence greater than optimal frequency. [9]

Schmalensee (1977) also derived similar results. He further identified circumstances where a form of destructive competition could emerge. This is if there are increasing returns to individual market shares with increased frequency, but decreasing returns to flight frequency for the market as a whole. Firms have an incentive to over invest because the market appears more attractive to each firm than it really is. Schmalensee, like Panzar (1979), finds that ideal regulation could improve upon the competitive market result.

The significance of these theoretical analyses are subject to debate. They assume a very enlightened and strong regulatory agency, and a costless one as well. Therefore, these models do not help much in determining: (1) whether or not regulation in practice can perform anything like the theoretical requirements; and (2) is the improvement in resource allocation worth the costs of regulation? These are important empirical questions which the theory does not address.

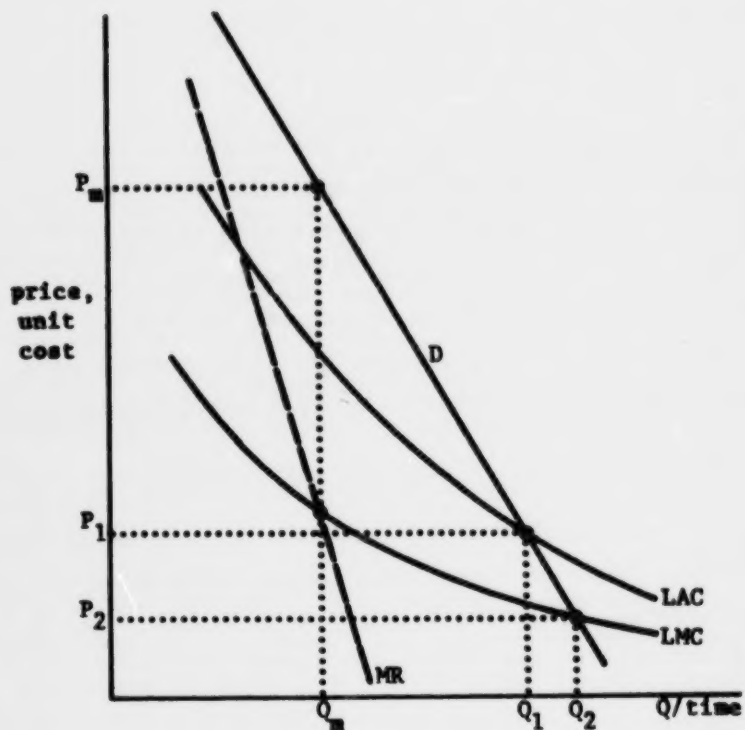
Nonetheless, these theoretical exercises do warn us against unqualified faith that competitive markets necessarily yield superior outcomes to what could be obtained with proper regulation.

D. Increasing Returns and Their Policy Implications

Increasing returns or decreasing unit costs have long been one of the basic rationales for regulation. [10] This is the so-called natural monopoly as pictured in Figure 1. The level of demand in the market is such that average costs decline for expansions in demand. We usually think of large scale enterprises such as electric power companies, communications, or traditionally, railways. But the natural monopoly concept applies on a small scale as well. The only scheduled truck or bus service to a small community could be an example.

Figure 1

NATURAL MONOPOLY AND THE DECREASING COST INDUSTRY PROBLEM



The concept of natural monopoly is important for regulation because demand is insufficient to support several competitors. Hence, regulation is necessary to prevent the monopolist from charging prices above marginal costs, or so the traditional argument goes. In Figure 1, the monopolist would charge price P_m for quantity Q_m . This has two implications for government policy: (1) there may be an income distribution concern because the monopolist may be getting wealthy at the expense of consumers; and (2) there is an inefficient allocation of resources because price is above marginal cost and buyers willing to pay prices above costs are not being supplied. The latter has been the chief concern of economic analysis of regulatory policies.

Ignoring second best considerations, the optimal policy is to use regulation to simulate the outcome of competitive markets, viz, price at marginal cost. As is well known, this gives rise to the "decreasing cost industry dilemma" that pricing at marginal costs results in a financial deficit. That is, the profit-making monopolist becomes financially insolvent. Average cost pricing enables the firm to break even, but the welfare dilemma remains that resources are not being allocated efficiently.

All this is well known, and still correct as far as it goes. But there have been several new insights into the nature of increasing returns and their implications. Some of these insights are independent of one another and others are related. The first topic discussed below can be discussed independent of the others; the latter are linked together.

1. Increasing Returns in User-Supplied Inputs [11]

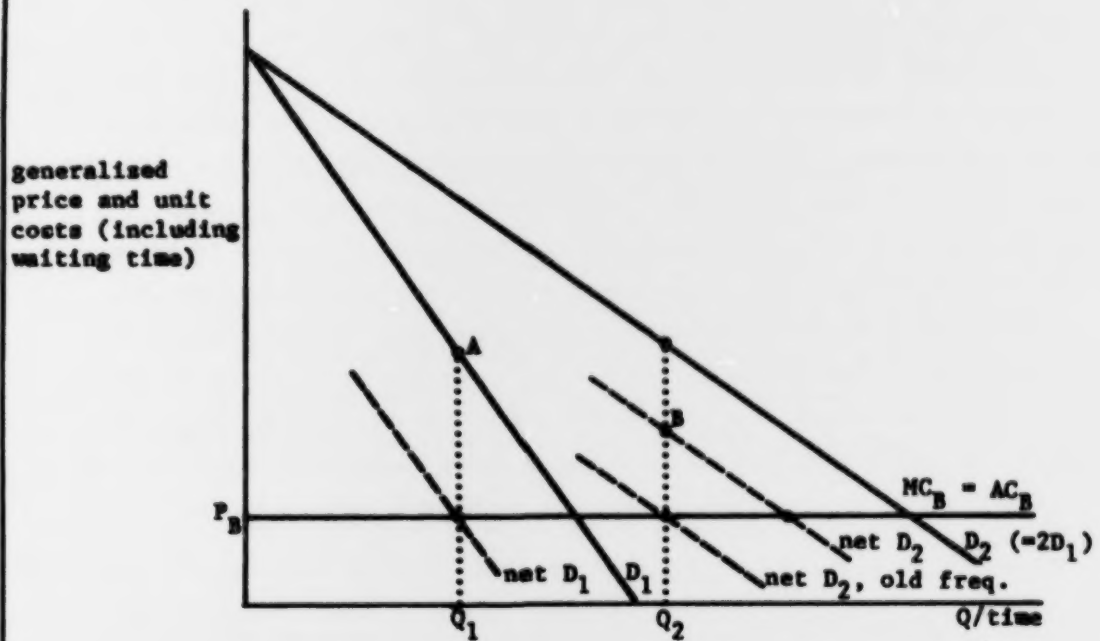
An important characteristic of transport is the dual role of transport users. They are both consumers of transport and producers in that they bear the costs of waiting time and other inconveniences associated with transport. The full costs of transport include both the costs borne by the producer of transport, but also the costs borne by the passenger or shipper of freight. Traditional discussions of increasing and decreasing returns have focussed on the producer's costs. But it is also possible, and in fact is likely, that increasing returns arise in user-supplied inputs.

Notice that the users do value the reduced waiting time. The reduction in the waiting time period point A to point B is directly reflected in an increased willingness to pay. But this is not the whole story. Users consider their own waiting time (and fare paid) in choosing to ride, indicated by points A and B in Figure 2. But because the collective result of increased bus use is increased frequency as the bus company expands operations, this implies that the change in total waiting time by all users as output expands is not the same as the average waiting time which is what each user considers. That is, the marginal social time cost curve must lie below the average time cost curve passing through points A and B. Each individual regards waiting time (bus frequency) as fixed. While the impact of one person on frequency is regarded as inconsequential, the collective result is significant. As usage increases, frequency will increase with the increased number of buses required. This external benefit of reduced waiting time for all riders is ignored by individuals. In order to bring about the social optimum, it is necessary to make bus users consider the true marginal costs of their decision rather than only their personal costs of time and money. The way to do this is to lower the bus fare by an amount equal to the difference in the average and marginal social time cost. The externality associated with the increasing returns is in time spent waiting, but the externality is corrected by adjusting the monetary outlay by individual riders.

This is illustrated in Figure 3 which is Figure 2 redrawn, showing the average and marginal social time cost. The optimum is where generalized price equals marginal social costs. For the original level of demand D_1 the optimal output is Q_1' instead of Q_1 . For demand D_2 the optimal output is Q_3' instead of Q^3 at price P^2 . Notice that the net willingness to pay (demand as perceived by the bus company) shifts outward as output expands reflecting the improved frequency with higher volumes. The optimal price for demand D_1 is P_1 . This is below the breakeven bus fare P_B (equal MC_B) by the difference between the marginal and average social time cost.

Figure 2

INCREASING RETURNS FROM REDUCED WAITING TIME OF BUS USERS



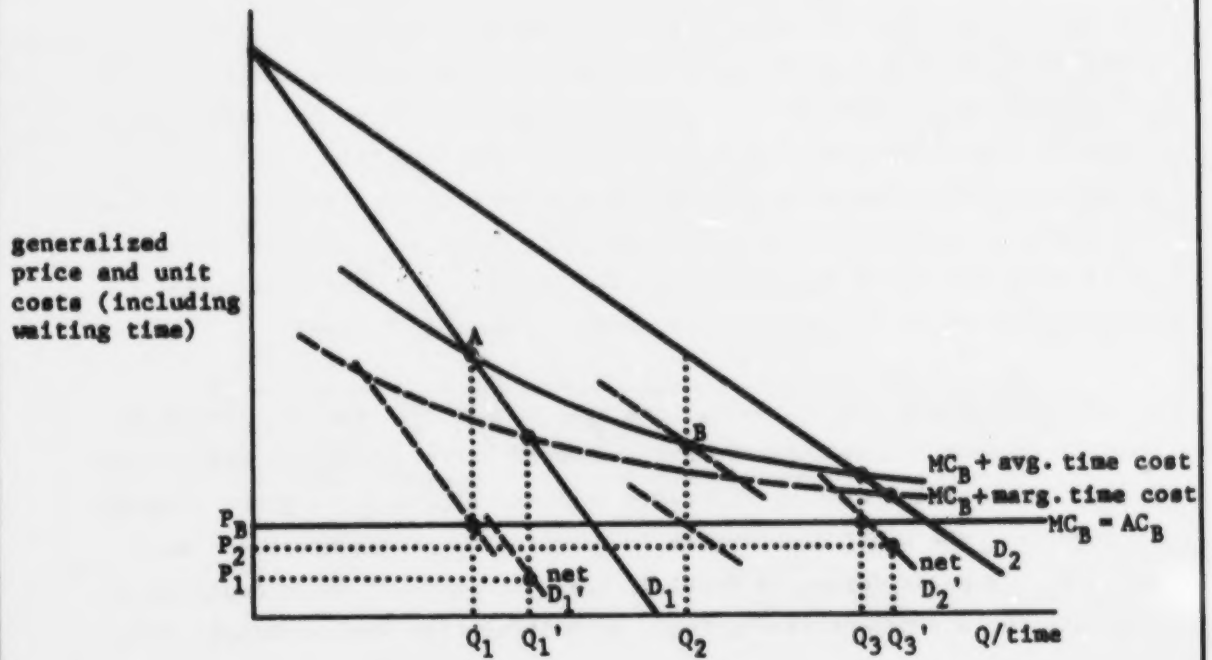
The demand for transport is continuous over time intervals whereas indivisibilities and/or economics of vehicle size result in supply at discrete time periods. Users of scheduled service encounter differences between their preferred versus available departure times. There is a waiting time delay due to the interval between departures/arrivals. [12] As traffic volumes grow, an increased number of vehicles are employed by the producer and this results in more frequent service. The reduced headway times result in reduced delay or waiting time to users. That is, the total costs of waiting time do not increase in proportion to producer costs as passenger volumes increase.

An example will illustrate the increasing returns in waiting time and the implications for pricing. Consider a generalized demand for bus services, i.e., the quantity of bus trips portrayed against total willingness to pay whether paid as monetary outlays or as personally-borne costs of time. Assume a linear and constant value of time. The total marginal social costs of supplying bus service are the costs incurred by the bus company, assumed to be constant per trip $MC_B = AC_B$, plus the costs of waiting time borne by the users. Bus size, travel time, and other characteristics of the transport system are assumed constant.

Figure 2 shows the initial generalized demand D_1 . The net demand or net willingness to pay curve is the demand as perceived by the bus company, i.e., the schedule relating ridership with fares paid for a given frequency of service. The usual equilibrium is quantity Q_1 at the breakeven bus price P_B . Suppose demand is doubled, i.e., a rightward shift doubling the quantity for any generalized price. If waiting time were unchanged (e.g., buses run two at a time rather than at intervals), the new net demand would intersect the bus cost curve at quantity Q_2 . [13] But running the buses at intervals reduces the headway time, hence reduces the waiting time of individual users (assuming random arrivals at bus stops). [14] The generalized costs of bus service per trip (bus cost plus value of waiting time) have decreased from the level at point A to the level at point B reflecting the increasing returns. At Q_2 the willingness to pay exceeds total unit costs (B) and it is optimal to expand.

Figure 3

OPTIMAL PRICES AND OUTPUTS UNDER INCREASING
RETURNS FROM REDUCED WAITING TIME



Source: Waters (1982)

Figure 3 shows that the average and marginal time costs converge to zero as waiting time approaches zero (bus movements become continuous). The more gentle slope as these curves approach zero reflects the diminishing returns from further increases in the number of vehicles (frequency). Therefore, the per unit subsidy (difference between average and marginal time costs) is larger for lower traffic volumes. This suggests that per unit subsidies would be greatest on less travelled routes, such as off peak and feeder routes rather than major corridors (Mohring, 1972; Jansson, 1980) and smaller bus sizes than those currently employed (Jansson, 1980).

Before leaving this topic I should stress that this discussion focussed only on waiting time. There are several other sources of delay and inconvenience costs borne by users which also involve externalities, some of them are negative such as congestion delays and crowding. The latter tend to be important on heavily travelled routes, i.e., those where the increasing returns just discussed are probably not so important. There are also possible increasing returns to producers, i.e., the traditional sources of decreasing costs. Thus, several factors are involved in determining optimal prices for scheduled transport services and it is not necessarily the case that the increasing returns will dominate. Optimal pricing could result in either a financial deficit or surplus. [15]

2. Decreasing Costs and the Multiproduct Enterprise

One of the hallmarks of transportation is that it consists of multiproduct enterprises, probably unmatched by any other industry. In spite of this importance, our formal understanding of the nature of multiproduct cost functions has been rather weak. Or, at least, economists only recently have been examining rigorously the properties of multiple product cost functions. The recent formulations (e.g., Baumol, et al.) have been of wide interest and several new insights have emerged. I think some of these findings are merely formulating what has been known intuitively or informally in transport economics for years. The terms "economies of scope" (Panzar and Willig, 1975 and 1981) and "transray convexity" (Baumol, 1977)

are no different from what transport economists have referred to for years as common costs. (However, transportation economists have used the term ambiguously. See my 1976 paper for a discussion.)

The proper phrase for identifying the multiproduct equivalent of a single product natural monopoly is a "subadditive cost function", i.e.,

$$C(X_1, X_2, \dots, X_m) < C(X_1) + C(X_2) + \dots + C(X_m)$$

That is, outputs can be produced at lower cost in combination with one another than if they were produced separately. But the multiproduct cost function can include very complex cross relationships among costs and various output combinations. "Strict subadditivity" is the analog to the single product natural monopoly. This is where there are increasing returns for equiproportionate expansions of any given combination of outputs and, for any given level of expenditure, more can be produced in common than the same expenditure could yield by producing the products separately. For any bundle of outputs the cost function would look like that of the traditional decreasing cost industry. This is shown in Figure 4. Baumol (1977) identifies this as "ray subadditivity" or "decreasing ray average costs" (Baumol, Bailey, and Willig, 1977). The other cost characteristic is the complementarity in production. Baumol identifies this as "transray convexity" and can be seen in Figure 5a. A linear combination of two (or more) outputs, costs less to produce together than if each were produced separately.

Figure 4
"RAY SUBADDITIVITY" OR "DECREASING RAY AVERAGE COSTS"

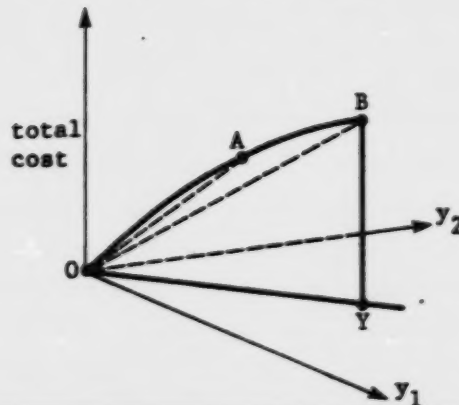
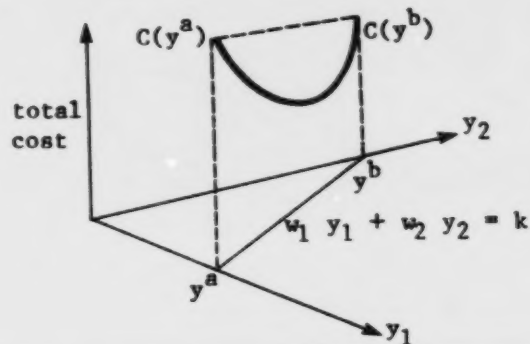


Figure 5a
TRANSRAY CONVEXITY OR ECONOMIES OF SCOPE



Expressed another way (Figure 5b), a given cost outlay can produce more of the outputs when produced together than if they were produced separately. (Historically, this is what transport economists meant by common costs.)

Figure 6 shows a strict subadditive cost function for two products. This is the multiproduct equivalent of the classic decreasing cost industry. Strict marginal cost pricing would lead to financial deficits just as for the single product case. Of course, the costing and pricing decisions are all the more complex to solve in the multiproduct case.

Figure 5b
IMPLICATIONS OF TRANSRAY CONVEXITY OR ECONOMIES
OF SCOPE (COMMON COSTS)

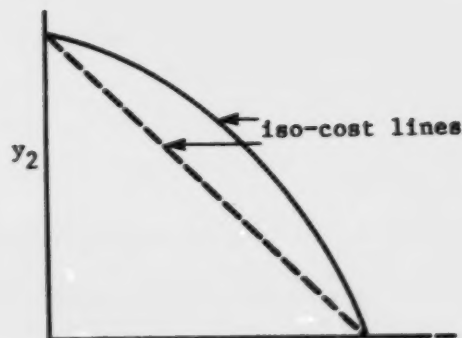
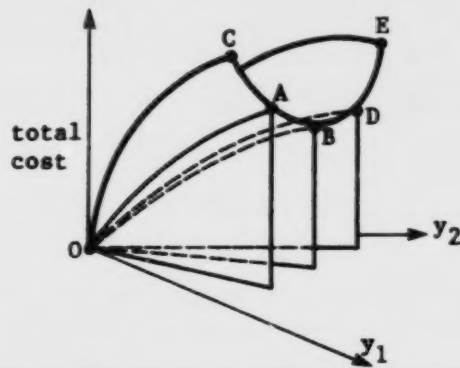


Figure 6
STRICT SUBADDITIVE COST FUNCTION



The monopolist would seek to charge higher prices than these; a regulatory agency would seek to prevent the monopolist from making monopoly profits. More on regulation and pricing shortly.

The foregoing is strict subadditivity. But the cost function can be much more complicated. With multiple products, scale economies may be more important for some products than others, or more important at different levels of output. Further, some combinations of multiple products may be complementary but others may conflict. Increasing the number of trains on a system may be more efficient than building a new track and facilities; at the same time, mixing fast and slow freight trains or passenger and freight operations may interfere with one another and may increase costs relative to a more homogeneous system. Further, the cost complementarities and scale effects could differ for different mixes of products and at different levels of output. It turns out that "...scale economies are neither necessary nor sufficient for monopoly to be the least costly form of productive organization." (Baumol, 1977, p. 809). Nor does complementarity in production (transray subadditivity) necessarily imply anything about scale economies.

The message to regulators is that natural monopoly conditions are not necessarily as obvious as might have been thought; and evidence of scale economies and cost complementarities for one level and a mix of outputs might not be a reliable guide to cost behaviour for different scales and

mixes of outputs. Hence, policies to encourage monopoly and regulate it as opposed to encouraging competition may require greater knowledge of cost conditions than were thought necessary in the past.

3. Sustainability, Contestability and Regulation

It is usually taken for granted that monopolists can make economic profits, and therefore need regulation. This assumes that there are barriers to entry which prevent others from entering the market and undercutting the monopolist. In recent years there has been much interest in the question of entry and its implications for regulation of monopoly, in particular, multiproduct monopoly. If entry were relatively easy in some or all markets served by a monopolist, then regulation may not be so necessary. This has given rise to the theory of contestable markets (Bailey, 1981; Baumol, 1982; and Baumol, Panzar and Willig, 1982). Where markets are contestable, regulation might not be necessary even if the market structure is monopolistic or oligopolistic in nature. On the other hand, there are also circumstances where entry is possible and profitable but it is not desirable from a social perspective, hence regulation may be necessary to restrict entry.

The concept of contestability replaces the emphasis on competition. Intuitively, we usually identify the effectiveness of competitive markets with the number of firms. Where the number is large, competition works; prices are equated to marginal costs and, at the same time, to minimum long run average costs. This is the textbook model of competitive markets. But where the number of firms is few or even one, the very different market structures of oligopoly or monopoly, respectively, prevail. We expect the outcome of these market structures to differ from that of perfect competition.

The theory of contestable markets focuses on entry and exit conditions. The logic is that if entry and exit are easy, then the threat of competitive entry is as good as competition itself. This can apply even if there are decreasing costs, i.e., in what is usually thought of as natural monopoly conditions. The threat of competitive entry may be sufficient to keep the

incumbent firm's prices down. (Indeed, paradoxically Baumol (1982, p. 14) points out that the absence of entry might be a sign of efficiency rather than market power as is traditionally thought; it could mean that the incumbent firm is keeping its price low enough to deter the easy entry.) In brief, the "invisible hand" which keeps markets efficient under perfect competition might also apply to cases of monopoly and oligopoly. It depends on entry and exit conditions. [16]

The contestability of markets is crucially related to the presence of sunk costs. More specifically, it is not increasing returns per se which generate a need for government regulation of a natural monopoly, it is where sunk costs are important (Bailey, 1981). The only bus or truck serving a small community is a natural monopoly if the optimal vehicle size is constrained by the small size of the market. Nonetheless, entry and competition is not wasteful because the investments can be relocated to other markets at minimal cost. The contestability argument is a rationale for allowing entry into airline markets despite the high cost of equipment. The aircraft can be moved to other markets if the firm cannot survive in a given market (e.g., Bailey, 1981). In contrast, if entry requires investing in major irretrievable facilities, then competition for a limited market is less feasible and is potentially wasteful. Irretrievable or sunk investments are a risky undertaking to a prospective entrant and the incumbent is in a position to undercut the price of the prospective entrant. The sunk costs act as a barrier to entry. This enables the existing firm to charge a higher price and make some monopoly profits. And if entry does take place, there is the wasteful duplication of facilities. Competition in these circumstances may not be desirable.

A further implication of the concept of contestable markets and sunk costs is that this can influence the design of regulatory regimes (Bailey, 1981). Sunk costs may be important in only part of the operations of multiproduct enterprises. By isolating the sunk cost portions for regulation and ensuring access to the sunk facilities for prospective entrants, this enables reliance on competition while retaining regulatory control over the relevant segment of the industry. A couple of examples come to mind. Provision of roads and airports involve sunk costs. These

involve the public sector in important ways but by permitting easy entry to prospective users, this results in a relatively competitive industry despite the presence of decreasing and sunk costs associated with the infrastructure. A similar argument could be applied to nationalizing railway track although there are many more pro's and con's on this matter which cannot be reviewed here. The central point is that contestable markets are those for which sunk costs are not important. In these cases competition, or simply the threat of entry, provide a restraint on the behaviour of a monopolist. But where sunk costs are important then markets are not so contestable and the case for government regulation is stronger. In these circumstances it may be possible to design the regulatory framework to focus on the sunk cost sectors and permit competitive access to the sunk cost components (cf., Bailey, 1981).

A major recent interest in the theory of multiproduct natural monopoly is "sustainability." A monopolist must be financially viable to survive. An unregulated monopolist would charge what the traffic will bear to maximize profits but is constrained by potential entry. If entry is easy this will severely limit potential economic profits. The alternative to market forces and entry is regulation. In either case it is desirable that the monopolist be financially viable. That is, the possibility of marginal cost pricing accompanied by public subsidy is ruled out. Ruling out the "first best" equilibrium (marginal cost pricing with a subsidy) is a fundamental feature of the sustainability analysis. Of course, a variety of prices can be set by the monopolist. Baumol, Panzar, Willig, and others have been focussing on the market conditions which limit the pricing freedom of a monopolist. To focus on the extreme they assume no barriers to entry. The question then is: is there a set of prices which enable the monopolist to breakeven without attracting entry into some of its product lines or other submarkets? This would be called a "sustainable" situation. [17] If a monopolist has a strict subadditive cost function, i.e., decreasing ray average costs and economies of scope in all product combinations, then the monopolist is sustainable despite freedom of entry.

The next question is what vector of prices is best for the monopolist and society? Some sort of average cost pricing is possible (although

"average costs" are ambiguous for a multiproduct firm). Of greater interest are the so called Ramsey prices. This is the set of prices where markups above costs reflect the willingness of different buyers to pay so that financial viability is achieved at minimum welfare loss associated with the deviation from marginal cost pricing.

4. Ramsey Pricing

A topic of considerable interest the last few years is that of Ramsey pricing, so named in honour of Frank Ramsey who first outlined the principles in 1927. The question of interest is: what is the most efficient (least economic welfare cost) way to raise a certain amount of revenue via a sales tax or equivalent markup over costs of production? More recently, it is Boiteux (1956, 1971) and, especially, Baumol and Bradford (1970) who are identified as elaborating the principles of pricing above marginal costs in order to satisfy a revenue constraint. The principles apply to any revenue constraint but the topic of greatest interest to transport economists is financing a decreasing cost industry. That is, a decreasing cost multiproduct monopoly would be permitted to charge prices sufficient to cover all costs rather than to receive a subsidy which would be necessary if the firm set prices at marginal costs. The basic principles are familiar to transport economists. Instead of the firm marking up prices by some uniform percentage, the optimal policy is to vary the markup according to the willingness to pay or "value of the service." Value of service pricing minimizes the welfare loss (total demand not served despite their willingness to pay exceeding marginal costs). If demands for various products are independent, the price markups vary inversely with the elasticity of demand for the product. The optimal price for each commodity i is given by:

$$\frac{P_i - MC_i}{P_i} = - \frac{B}{e_i}$$

where e_i is the price elasticity of demand, B equal to zero corresponds to perfectly competitive or "first best" prices and B equal to unity corresponds to the unrestricted monopoly case, i.e., profit maximization. B is set between zero and one depending on the revenue to be generated.

The Ramsey pricing formula is much more complex if the cross elasticities of demand are not zero as is assumed in the above expression. A full matrix of elasticities and cross-elasticities must be incorporated when solving for the optimal price markups.

It turns out that there is a fairly simple rule to guide optimal pricing in both the simple and more complex cross elasticity case. The "optimal departures from marginal cost pricing" or "quasi-optimal prices" (Baumol and Bradford, 1970) are those which yield an equiproportionate reduction in output from those corresponding to $P = MC$. Hence, lower markups are in order for more elastic demands taking into account cross elasticities. The rule is simple conceptually although it would not be so simple to implement in practice.

The Ramsey pricing argument is of particular interest and relevance to Canadian transportation. Stated briefly: given that railways have overheads and other sources of decreasing costs (including uncompensated losses) which must be covered, and that we wish railways to be self financing, i.e., not relying on public subsidy for their operations, then the economically optimal policy is to allow value of service pricing up to the point where the railways begin to earn more than competitive rates of return. This is an economic rationale for the pricing freedom given to Canadian railways.

Two issues grow out of this which must be mentioned. First, the argument for free reign of value of service pricing is valid until the railways' rate of return reaches competitive levels, i.e., the financial goal is achieved. If railways begin to earn economic profit then regulation is necessary for classic monopoly reasons. Note that if earnings do become excessive, the optimal regulatory policy is not to reduce rates by uniform amounts. Optimal taxation policy (Ramsey pricing) calls for differential price cuts reflecting the different elasticities of demand for transport. The optimal rule to decrease rail profits would be to require rates to be cut such that there would be an equiproportionate increase in all traffic

relative to the levels which would exist in the absence of the price reductions. As mentioned, this rule is simple in principle but would be difficult to implement in practice.

A second point about Ramsey pricing and its applicability to Canadian railways is that the argument thus far is valid as applied to a single sector of the economy, the railways in this case. But the Ramsey pricing principles can be applied across more than one sector for greater efficiency gains. I return to this point in Section 6.

5. Ramsey Pricing and Sustainability

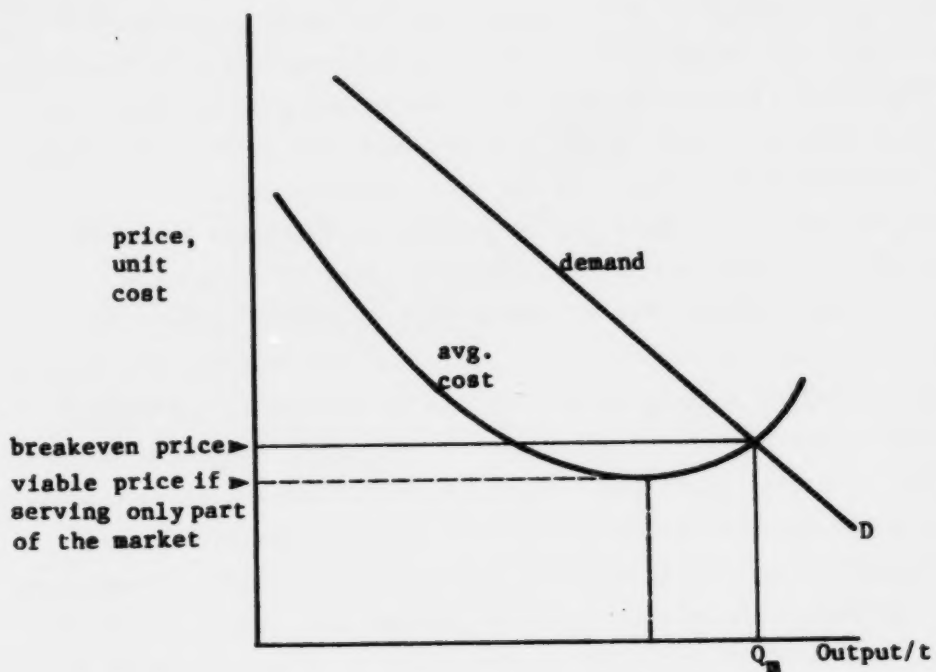
Baumol, Bailey and Willig (1977) prove a "weak invisible hand" theorem which shows that where entry is costless, a multiproduct natural monopoly is held check by the threat of entry. In fact, Ramsey prices are sustainable with zero economic profits for the firm with a strict subadditive cost function. Thus, the invisible hand not only applies to competitive markets. Government intervention may not be necessary even for natural monopolies if its markets are contestable (free entry).

The crucial conditions are two. First, is freedom of entry which serves as a competitive check on the pricing behaviour of the monopolist. Second, cost conditions must be such that the monopolist is sustainable. Both of these require elaboration, the latter is taken first.

If the cost function contains some increasing cost elements and/or transray concavity (some product combinations cost more in combination than if produced separately), then it is possible that, with free entry, others could supply submarkets profitably at prices below those of the monopolist. This could undermine the sustainability of the monopolist. This is primarily a multiproduct phenomenon but I can give two simple illustrations of this possibility. The first is Figure 7 taken from Panzar and Willig, 1977. If there are increasing costs over a range of output, a subgroup of consumers could break off and be supplied profitably at a price below that of the monopolist even if it were only breaking even as shown in Figure 7. However, the regulated monopolist is obliged to serve all traffic at the

Figure 7

A SUSTAINABILITY PROBLEM WITH FREE ENTRY



Source: Panzar & Willig (1977)

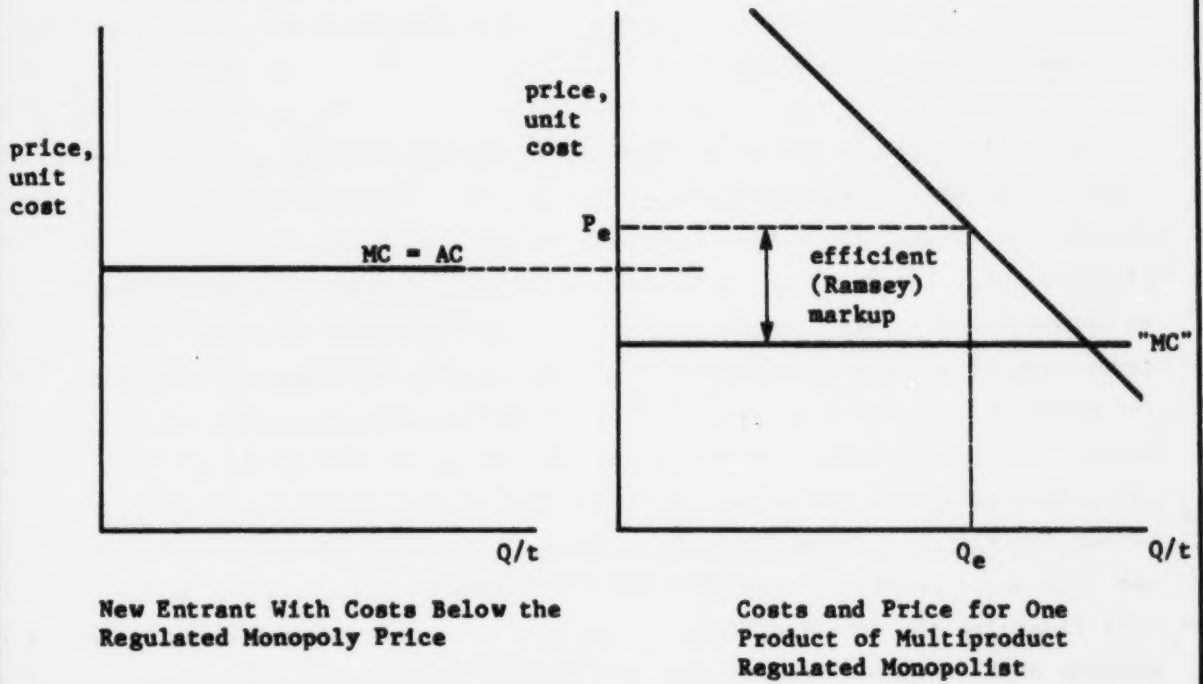
break-even price. A new entrant would only seek to serve a limited portion of the market, viz., the traffic which can be served at minimum average cost and leave the higher cost traffic to the incumbent firm. But then the original firm is no longer sustainable. This is very reminiscent of the old "cream skimming" argument in transportation -- and this shows there can be some logic to the argument. The crucial assumptions are that there is only room for one firm and increasing costs are encountered at the margin. [18]

A different illustration of the sustainability problem is Figure 8. Keep in mind that the sustainability of the multiproduct natural monopolist requires that total revenues be covered, hence prices will be above the marginal costs and, if Ramsey prices are employed, the markups will vary according to willingness to pay. If the cost function is not strictly subadditive, or barely so, there may be submarkets which can be profitably invaded by new suppliers despite having higher costs of production than the monopolist. Figure 8 shows that the monopolist charges a price slightly above marginal costs ("MC" in quotes to remind us that this is the marginal cost in a multiproduct enterprise and "MC" is determined in conjunction with all the other operations of the enterprise). The price-cost divergence is necessary for financial viability. Figure 8 shows that the entrant would be profitable at the monopolist's price but the costs of the entrant are higher. Society is not better off with the entrant and such circumstances threaten to undermine the viability of the monopolist. Hence, it may be optimal for regulation to not allow a competitor to enter and charge a price that was below the price charged by the monopolist. [19]

So, we see situations where free entry may not be desirable given elements of decreasing costs in existing firms. The crucial underlying assumptions are that (1) firms with decreasing costs are required to be self-sustaining financially, and (2) an efficient regulatory regime is in place. The assumption is that with regulation we can achieve the efficient (Ramsey) prices necessary to raise the revenue to break even in the decreasing cost firm. [20] Ramsey pricing is sustainable for financing the multiple product natural monopoly if the cost function is strictly subadditive. But there is no reason to think that cost functions are strictly subadditive, that is,

Figure 8

ILLUSTRATION OF ENTRY CONTROL TO PRESERVE THE
SUSTAINABILITY OF A REGULATED MONOPOLY



there are likely to be increasing cost elements in some of the firm's operations. This is certainly the case for railways. Hence, entry to rail markets may interfere with efficient Ramsey pricing. Still more specifically, it will be seen that the presence of intermodal competition raises serious issues about the feasibility and desirability of Ramsey pricing.

6. Ramsey Pricing and Intermodal Competition

There are two issues here. The first, is the "second best" principle that charging a price above marginal costs in one mode should entail price changes elsewhere to maintain an efficient allocation of resources among sectors of the economy. The second point is that revenue constraints (hence the need for Ramsey prices) should apply across modes (or sectors of the economy), i.e., optimal taxation should apply across modes of transport and not just be levied on a single mode.

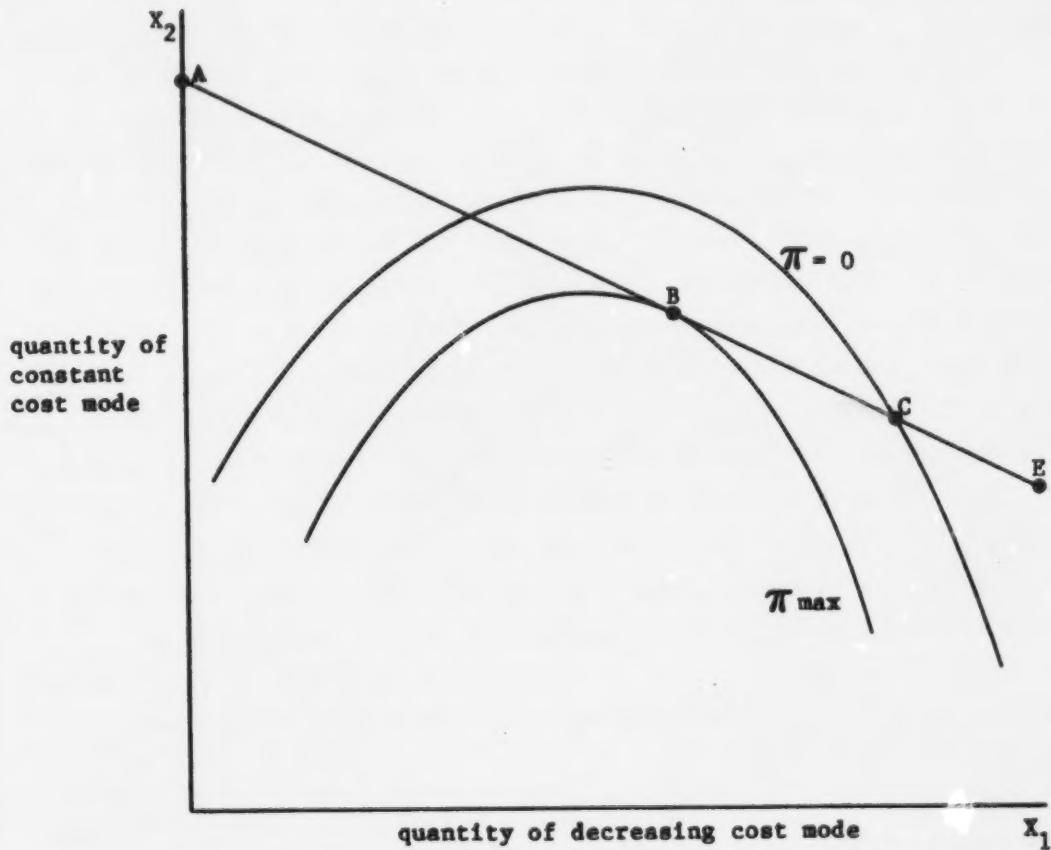
An illustration of the first point about optimal pricing and intermodal competition comes from Braeutigam, 1979. He has a formal mathematical analysis but I will limit the discussion to his simple geometric illustration. Assume there is one mode of transport producing quantities X_2 at constant costs, e.g., motor carriers. X_1 is the output of a mode with decreasing costs, e.g., railways. There are service differentials between the modes so there is a range of prices and market shares between the modes. In Figure 9, the locus of points AE are quantities of X_1 and X_2 corresponding to various prices of X_1 and the resulting market clearing prices for X_2 . [21] Point E is the optimum of marginal cost pricing for X_1 , but this would require a subsidy for X_1 . If mode 1 is constrained to be self financing the "quasi-optimal" price will be above marginal costs. The amounts of X_1 supplied will decline and that of X_2 will increase. Some further analysis is in order to explain this move and a more efficient alternative.

Looking at X_1 , a given level of profits could be obtained from different prices (and quantities) of X_1 and X_2 although only the quantities given by the locus AE are market clearing prices. Two iso-profit lines are shown in

Figure 9

LOCUS OF EQUILIBRIUM QUANTITIES
OF X_1 (DECREASING COST MODE) AND
 X_2 (CONSTANT COST MODE), FOR VARIOUS
PRICES OF X_1

Iso-profit lines for X_1



Source: Braeutigam (1979)

Figure 9. The profit maximizing point is B. (The intersections of iso-profit contours to the left of B are not of interest). A breakeven constraint (zero profit) for X1 results in point C. Braeutigam (1979) refers to this as "Partially Regulated Second Best," i.e., the price of the decreasing cost mode is allowed to rise to make it self financing and the constant cost mode X2 is allowed to supply whatever it can at the price of X1.

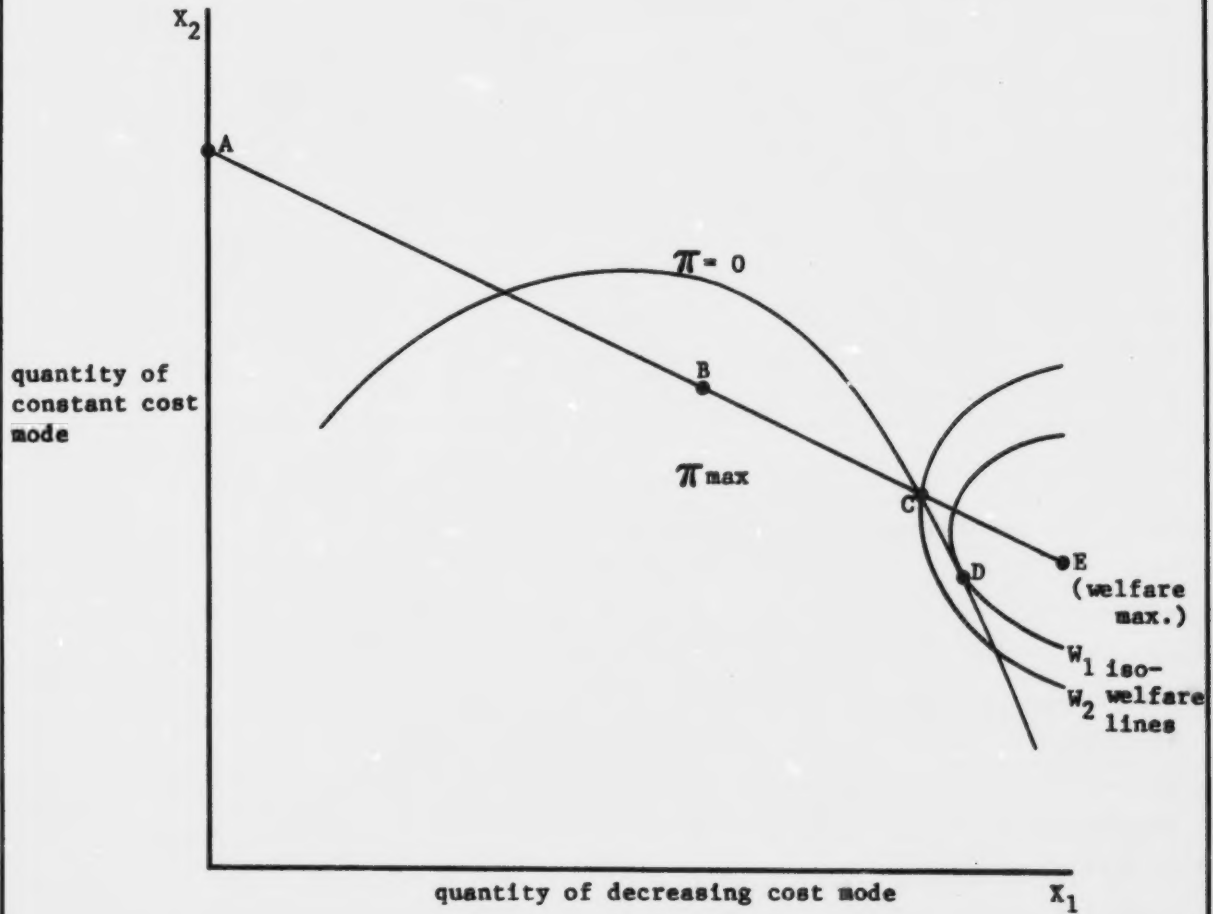
But there is a more efficient solution. The theory of second best tells us that once a constraint moves us away from marginal cost pricing in one sector of the economy, it is optimal for price to exceed marginal cost in a substitute sector. This can be illustrated as follows. A welfare function (sum of producers' and consumers' surplus) can be visualized above X1 and X2. It is a mound with a maximum above point E. We can project iso-welfare lines on to the X1-X2 diagram. This is shown in Figure 10. The iso-welfare lines are vertical where they intersect AE. This is because the market clearing price for X2 is the most efficient price (and quantity) given the price for X1 (see Braeutigam, 1979, for the mathematical proof). The constrained welfare maximum is the tangency between the zero profit curve for X1 (its breakeven constraint) and an iso-welfare curve. This is point D in Figure 10. The amount of X2 is less than on AE indicating a higher price of X2. This is not a market clearing price for X2, hence government intervention would be necessary to sustain the price. This could be accomplished by selective restrictive entry to raise prices in that mode or by levying a tax on X2. Point D is what Braeutigam calls "Total Regulated Second Best."

The foregoing shows that second best pricing principles call for intervention in competitive markets if we wish the most efficient allocation of resources but subject to the constraint of making the decreasing cost sector self-financing. But whether the welfare gains moving from C to D are worth the administrative costs is an empirical question, and a very important one which the theory does not answer.

I now wish to take this analysis of optimality and intermodal competition one more step. We usually think of the motor carrier industry

Figure 10

SECOND BEST PRICES FOR DECREASING
COST MODE CONSTRAINED TO BREAK-EVEN



Source: Braeutigam (1979)

as a constant cost industry. And it is from the point of view of the carriers. But that is because the infrastructure costs are borne by the public and a variable charge imposed for the use of roads. Congestion is not normally a problem for intercity transportation; hence there is a good deal of underutilized capacity. In brief, road infrastructure is a decreasing cost sector yet we impose prices for use in excess of marginal costs. But the taxation principle applied to financing roads is different from the implicit taxation principle (price markup according to the value of service) applied to railways. User charges may vary with the size of road vehicle but they do not vary according to what is in the vehicle. The motor carriers may be able to engage some value of service pricing, particularly if regulatory restrictions on entry are present, but all of us would agree that the opportunities for Ramsey type value of service pricing are more limited in trucking than in the railways. In brief, the implied taxation principles for recouping the infrastructure costs in road and rail are different. One can appreciate that there are potential welfare gains if a consistent taxation-pricing policy could be applied across modes.

One of the principles from Boiteux (1956, 1971) is that the optimal taxation framework should be applied as widely as possible. That is, rather than set taxes (price markups) for one sector of the economy, the revenue requirements and tax principles should be coordinated across sectors of the economy. Value of service pricing by railways is consistent with satisfying the constraint that the railways be financially viable despite the presence of decreasing cost operations and possible unremunerative public obligations. But if one groups other modes together and considers overall revenue needs for all modes and sets an optimal tax policy to raise this revenue, the taxes or price markups will probably differ from those resulting solely from rail value of service pricing. Optimal tax policy would entail different markups according to the willingness to pay for transport regardless of mode. But where intermodal competition is present, the market prevents rail value of service pricing from exploiting the willingness to pay by this traffic. Intermodal competition undermines the sustainability of multimodal Ramsey pricing. Markups are below the optimum for mode competitive traffic and, because the rail revenue constraint is unchanged, markups are higher than optimal for non-competitive traffic and,

because the rail revenue constraint is unchanged, markups are higher than optimal for non-competitive traffic. But to implement an optimal pricing-taxation policy across transport modes would require massive regulatory intervention. It would mean that responsibility for pricing would rest with the regulatory agency (or it requires public monopoly of all transport modes which, in this case, is no different).

Where does this theoretical exercise lead? It leads to an argument that extensive government intervention is necessary if we wish to achieve an optimal allocation of resources in transportation along with a requirement that any decreasing cost components (and any other revenue requirements in excess of marginal costs) be financed from transportation users rather than by public subsidy. It presumes that the goal of regulation is economic efficiency subject to revenue constraints, and it presumes an efficient, costless and all-wise government regulation. No offense intended but....

Taplin and Waters (1981) have made some illustrative calculations of the impact of a Ramsey-Boiteux optimal pricing framework for Australia. The analysis was limited to road and rail freight transportation. Commodities were aggregated to four major groups. Cost functions for road and rail transportation were constructed along with estimates of own and cross price elasticities of demand. Then the quasi-optimal prices were estimated to generate the same government revenue surplus presently accruing from these modes (railways more than cover their costs on freight but incur substantial losses on passenger operations). The resulting price markups or taxes entail a significant shift of traffic from road to rail and a significant shift in the taxation burden from rail non-competitive traffic to mode competitive traffic. [22]

E. Conclusion

The economic principles discussed here lead us toward greater rather than less government intervention in transportation markets. However, this is a theoretical exercise. The theory seems to lead to either monopoly in all transportation and subject to detailed regulation, or the equivalent in this case, government monopoly in transportation. The usefulness of the

theory depends on the feasibility and costs of achieving more efficient pricing via government intervention as compared with the benefits of improved resource allocation. Many economists are still sceptical of regulatory performance in terms of economic efficiency hence they may question the relevance of the foregoing analysis (e.g., Shepherd, 1981). But the principles are there. Given the constraint that the transport sector is to be self-financing despite the presence of some decreasing costs, there is a case for government intervention to help achieve an optimal price structure where competition would otherwise distort it.

The next logical step is to question why single out the transportation sector? A more comprehensive and efficient policy is to integrate optimal taxation across still more sectors. We are being led down the path toward the Lange-Lerner type socialism. In theory, perfect government could improve upon the economy given the market complications and constraints I have mentioned. But should we follow this line of reasoning in practice? Just as real markets differ significantly from perfect textbook ones, so real governments differ from textbook ideals. The information task is astronomical, infeasible for all practical purposes. The theory assumes that a central authority can know cost and demand conditions as well as or better than individual firms in the market place. Like all comparative static welfare economics, it ignores potential dynamic effects of market competition such as the incentive to innovate. These are dubious propositions hence we should be cautious in embracing the various arguments explained here. In theory, the desire to make decreasing cost sectors pay their way and a concern to do this via optimal price markups, generates a logical case for wider government intervention in pricing and entry in transport markets, or any other market for that matter. But, as mentioned, this is predicated on an all-seeing and costless government which needs to intervene in order to impose the optimal tax structure. All this is a far cry from the more pressing need for economic advice and that is whether or not one regulatory (or promotional) policy will improve or worsen the economic performance of the economy and transport sector which exist, rather than the world of theory. My personal convictions lie strongly toward

reliance on the market, but the recent developments in economic theory relevant to regulation which I have been discussing provide little guidance here.

This review is, unavoidably, incomplete and sketchy. But I think it shows that there are a number of interesting developments taking place in economic thought, many of which are potentially important for the design and implementation of transport regulation. Just how relevant and important some of these ideas are remains to be seen. It is ironic that some economic research is beginning to identify situations where government regulation may be desirable. This is in stark contrast to the majority of economic analysis of regulation which has concluded that the economic performance of regulatory institutions has not been encouraging. But as a colleague pointed out when I brought one of these points to his attention: the economic criticisms of regulation show that regulations conceived in the early 20th century or during the depression are not efficient in the late 20th century. This does not mean that there isn't some late 20th century role of regulation which would be superior to no regulation at all. Whether or not these recent developments point the way remains to be seen, but they are avenues which merit attention.

reliance on the market, but the recent developments in economic theory relevant to regulation which I have been discussing provide little guidance here.

This review is, unavoidably, incomplete and sketchy. But I think it shows that there are a number of interesting developments taking place in economic thought, many of which are potentially important for the design and implementation of transport regulation. Just how relevant and important some of these ideas are remains to be seen. It is ironic that some economic research is beginning to identify situations where government regulation may be desirable. This is in stark contrast to the majority of economic analysis of regulation which has concluded that the economic performance of regulatory institutions has not been encouraging. But as a colleague pointed out when I brought one of these points to his attention: the economic criticisms of regulation show that regulations conceived in the early 20th century or during the depression are not efficient in the late 20th century. This does not mean that there isn't some late 20th century role of regulation which would be superior to no regulation at all. Whether or not these recent developments point the way remains to be seen, but they are avenues which merit attention.

FOOTNOTES

- [1] Roy (1980) and Caves and Christensen (1978). The translog function was first introduced by Christensen, Jorgenson and Lau (1973). There have been several applications of the translog function to transportation industries particularly by researchers at the University of Wisconsin and M.I.T. See Waters and Woodland (1982) for a summary of the application to the railway sector.
- [2] It is ironic that some of the most sophisticated social benefit cost studies are carried out in so-called third world countries rather than in the industrialized nations. This reflects the extensive use and emphasis on such studies by the World Bank and the United Nations. Detailed project manuals on project appraisal have been prepared. See Treasury Board (1976) for a brief summary of social benefit cost principles in a Canadian context.
- [3] The references are too numerous to mention. At a minimum I should cite Walters (1961), Johnson (1964), the Smeed Report (1964). Research continues to appear on this topic.
- [4] Important references on optimal investment strategies without socially optimal pricing would be Wheaton (1978) and Friedlaender (1981). The basic implications of non-optimal pricing for investment decisions have been recognized in England for some years, e.g., Gwilliam and Mackie (1975) 209-210 for a brief discussion.
- [5] The original formulation of the theory of second best is Lipsey and Lancaster (1956-57). There is extensive literature following from the original contribution. See Turvey (1971) for a pragmatic framework for evaluating second best prices.
- [6] Again, references are too numerous to review. In addition to the original formulation by Averch and Johnson (1962), see Bailey (1973), Baumol and Klevorick (1970) and Zajac (1970).
- [7] The number of authors are numerous. One might begin by citing Meyer, et al. 1959, but there were many previous authors who identified adverse economic consequences of regulation. On attempts to establish a dollar value to the costs of inefficient regulation see Moore (1972); also Friedlaender (1969 and 1971), Boyer (1977), Winston (1981).
- [8] This is an important assumption. It is well known that competition under differentiated products may lead to minimum differentiation between them. Suppose air passengers wish to depart between 7:00 and 9:00 in the morning and desired departure times are uniformly spread between these hours. Two airlines competing would tend to both schedule departure at 8:00 whereas the greatest convenience to passengers

passengers would have the two aircrafts scheduled at 7:30 and 8:30. The absence of this "locational rivalry" is an important assumption in his analysis.

- [9] As seen in the first part of the next section, there can be increasing returns to users with increased frequency of service, particularly on low density routes. But these increasing returns take the form of an externality. Since competitive markets do not incorporate the externality, this implies that competitive markets may undersupply service to low density routes.
- [10] For a thorough discussion of the concept of natural monopoly and its regulation see Posner, 1969.
- [11] This section draws heavily from Waters, 1982.
- [12] This is known as "frequency delay" in the literature on airline pricing and economics, e.g., Douglas and Miller, 1974. There are also queueing delays or "stochastic delays" (Ibid.) when vehicles are full and passengers or shippers must wait for the next departure. See Jansson, 1979, for a more complete discussion of categories of time costs borne by users.
- [13] It is possible that a bus company could run buses twice the size at the same frequency. Economies of vehicle size are a source of increasing returns to the bus company. These potential savings must be compared with the value of time saved by increased frequency to see which is the socially optimal policy.
- [14] In the special case where the bus company expands, buses in direct proportion to the increased demand (e.g., doubling demand would call forth double the supply of buses) the total waiting time is constant and marginal time costs are zero. The relevant marginal social cost curve is simply $MC_B (=AC_B)$ and an even greater subsidy is called for than shown in Figure 3. Unfortunately, this point was not incorporated in my forthcoming (1982) paper.
- [15] See Turvey and Mohring, 1975 for a concise discussion of these various cost concepts and their implications for optimal pricing. See also Jansson, 1979.
- [16] The importance of the threat of entry as an effective competitive force has been recognized by many authors of years. Nonetheless, this has been given only limited attention in formal theoretical analysis.
- [17] Sustainability is an issue in many of the references cited in this review. In particular see Panzar and Willig, 1977 and Zajac, 1978, Chapter 7.
- [18] One of the criticisms of sustainability analysis (e.g., Shepherd, 1981) is that it does not allow the incumbent firm to respond to the price offered by the new entrant (perhaps because of restrictive regulations?). However, this might not be a serious criticism. The

point remains that a subset of the market could attract entry at a price below what is necessary for the financial viability of the industry as a whole.

- [19] See also Faulhaber, 1975, on Ramsey pricing frameworks in contrast to the situation where prices are limited by competitive entry.
- [20] See Shepherd, 1982, for a critique of the relevance of the sustainability arguments for actual regulatory policy.
- [21] A limitation of the geometric illustration is that the existence of Ramsey price structures within a mode is hidden from view. If the decreasing cost mode was a single product firm, the breakeven price would have to be average costs. One can construct a locus like AE for the multiproduct firm where prices do not conform to Ramsey Prices, e.g., the multiproduct equivalent of average cost pricing. In that case the locus of market equilibrium quantities AE would not be socially optimal.
- [22] In terms of Braeutigam's geometry, applying the revenue constraint to both modes à la Boiteux means that tax revenue collected from the constant cost mode would reduce the revenue requirement for the decreasing cost sector. Hence, the totally regulated second best prices would be at the tangency of an iso-welfare line with the relevant profit locus where profits were negative but not as low as with pure marginal cost pricing. The optimum would be to the right of point D in Figure 10.



BIBLIOGRAPHY

- Averch, H., and Johnson, L. L. "Behavior of the Firm Under a Regulatory Constraint." American Economic Review (Dec. 1962), 1053-1069.
- Bailey, Elizabeth E. Economic Theory of Regulatory Constraint, Heath-Lexington, 1973.
- Bailey, Elizabeth E. "Contestability and the Design of Regulatory and Antitrust Policy." Amer. Econ. Review (May 1981), 179-183.
- Baumol, W. J. "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry." Amer. Econ. Review (Dec. 1977), 809-822.
- Baumol, W. J. "Contestable Markets: An Uprising in the Theory of Industry Structure." Amer. Econ. Review (Mar. 1982), 1-15.
- Baumol, W. J.; Bailey, E. E.; and Willig, R. D. "Weak Invisible Hand Theorems on the Sustainability of Natural Monopoly." Amer. Econ. Review (June 1977), 350-65.
- Baumol, W. J., and Bradford, D. F. "Optimal Departures from Marginal Cost Pricing." Amer. Econ. Review (June 1970), 265-283.
- Baumol, W. J., and Klevorick, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." Bell Journal of Economic and Management Science (Autumn 1970), 162-190.
- Baumol, W. J.; Panzar, J. C.; and Willig, R. D. Contestable Markets and the Theory of Industry Structure. San Diego: Harcourt Brace Jovanovich, (1982).
- Boiteux, M. "Sur la gestion des Monopoles publics astreints à l'équilibre budgétaire." Econometrica, (Jan. 1956), translated as "On the Management of Public Monopolies Subject to Budgetary Constraints." Journal of Economic Theory Vol. 3, (1971).
- Borins, Sanford F. "Pricing and Investment in a Transportation Network: The Case of Toronto Airport." The Canadian Journal of Economics (Nov. 1978), 680-700.
- Boyer, K. D. "Minimum Rate Regulation, Modal Split Sensitivities, and the Railroad Problem," Journal of Political Economy (June 1977), 493-512.
- Braeutigam, R. R. "Optimal Pricing with Intermodal Competition." Amer. Econ. Review (Mar. 1979), 38-49.
- Carlin, A., and Park, R. E. "Marginal Cost Pricing of Airport Runway Capacity." Amer. Econ. Review (June 1970), 310-319.
- Caves, D. W., and Christensen, L. R. "Productivity in Canadian Railroads 1956-75." Research Report 78-16E, Canadian Transport Commission, (1978).

- Caves, L. R.; Jorgenson, D. W.; and Lau, L. F. "Transcendental Logarithmic Production Frontiers." Rev. of Econ. and Statistics (Feb. 1973), 28-45.
- DeVany, A. S. "The Effect of Price and Entry Regulation on Airline Output, Capacity and Efficiency." Bell J. of Econ. (Spring 1975), 327-345.
- Douglas, G. W. and Miller, J. C., III. Economic Regulation of Domestic Air Transport: Theory and Policy. Washington, D.C.: The Brookings Institute 1974.
- Douglas, G. W., and Miller, J. C., III. "Quality Competition, Industry Equilibrium and Efficiency in the Price Constrained Airline Market." Amer. Econ. Review (Sept. 1974), 657-669.
- Faulhaber, G. R. "Cross-Subsidization: Pricing in Public Enterprise." Amer. Econ. Review (Dec. 1975).
- Friedlaender, A. F. The Dilemma of Freight Transport Regulation. Washington, D.C.: The Brookings Institute 1969.
- Friedlaender, A. F. "The Social Cost of Railroad Regulation." Amer. Econ. Review (May 1971).
- Friedlaender, A. F. "Price Distortions and Second Best Investment Rules in the Transportation Industries." Amer. Econ. Review (May 1981), 389-393.
- Gwilliam, K. M., and MacKie, P. J. Economics and Transport Policy. London: George Allen and Unwin Ltd., (1975).
- Jansson, J. O. "Marginal Cost Pricing of Scheduled Transport Services." Journal of Transport Econ. and Policy (Sept. 1979), 268-294.
- Johnson, M. Bruce. "On the Economics of Road Congestion." Econometrica (Jan. - Apr. 1964).
- Jordan, William A. Airline Regulation in America: Effects and Imperfections. Johns Hopkins Press 1970.
- Levin, Richard C. "Railroad Rates, Profitability Welfare Under Deregulation." Bell J. of Econ. (Spring 1981).
- Levin, Richard C. "Railroad Regulation, Deregulation, and Workable Competition." Amer. Econ. Review (May 1981), 394-398.
- Lipsey, Richard G., and Lancaster, K. "The General; Theory of Second Best." Rev. of Econ. Studies (1956-57).
- Meyer, J. R., et al. The Economics of Competition in the Transportation Industries. Harvard University Press, (1959).
- Mohring, H. "Optimisation and Scale Economies in Urban Bus Transportation." Amer. Econ. Review (Sept. 1972), 591-604.

- Moore, T. G. Freight Transportation Regulation: Surface Freight and the Interstate Commerce Commission. Amer. Enterprise Institute 1972.
- Panzar, J.C. "Equilibrium and Welfare in Unregulated Airline Markets." Amer. Econ. Review (May 1979), 92-95.
- Panzar, J. C., and Willig, R. D. "Economies of Scale and Economies of Scope in Multi-Output Production." Econ. Disc. Paper No. 33, Bell Laboratories, (1975).
- Panzar, J. C., and Willig, R. D. "Free Entry and the Sustainability of Natural Monopoly." Bell J. of Econ. (Spring 1977), 1-22.
- Panzar, John C., and Willig, R.D. "Economies of Scope." Amer. Econ. Review (May 1981), 268-272.
- Posner, R. A. "Natural Monopoly and Regulation." Stanford Law Review (Feb. 1969), 548-563.
- Posner, R.A. "Theories of Economics Regulation." Bell Journal of Economic and Management Science (Autumn 1974), 335.
- Ramsey, Frank. "A Contribution to the Theory of Taxation." Econ. J. (Mar. 1927), 47-61.
- Roy, Roger. "Economies of Scale in the Airline Industry." Research Report No. 10-80-06, Canadian Transport Commission, (1980).
- Schmalensee, R. "Comparative Static Properties of Regulated Airline Oligopolies." Bell J. of Econ. (Autumn 1977), 565-576.
- Sharp, Clifford H. Transport Economics. London: Macmillan, (1973).
- Shepherd, W.G. "Public Enterprises" Purposes and Performance, and Strategy and Structure," Paper for a Conference on Managing Public Enterprises: Purposes and Performance. Vancouver, Canada, (Aug. 1981). Proceedings forthcoming: Institute for Research on Public Policy.
- Smeed, R. J., et al. Road Pricing: The Economic and Technical Possibilities. Ministry of Transport HMSO, (1964).
- Stigler, G. "The Economic Theory of Regulation." Bell Journal of Economic and Management Science (Spring 1971), 3-21.
- Treasury Board Secretariat, Planning Branch. Benefit-Cost Analysis Guide. Supply and Services Canada, Ottawa, (1976).
- Turvey, R. Economic Analysis and Public Enterprise. Allen and Unwin: London, (1971).
- Turvey, R., and Mohring, H. "Optimal Bus Fares." Journal of Transport Economics and Policy (Sept. 1975), 280-286.

- Vanag, A. H. "Maritime Congestion: An Economic Analysis," In R. O. Goss, ed. Advances in Maritime Economics, Cambridge Univ. Press, (1977).
- Walters, A. A. "The Theory and Measure of Private and Social Cost of Highway Congestion." Econometrica (1961), 676-699.
- Waters, W. G., II. "Optimal Pricing Under Joint and Common Costs: A Geometric Analysis." Working Paper, Transport Studies Unit, Oxford University, (1976).
- Waters, W. G., II. "Waiting Time and Increasing Returns in Scheduled Transport Services: A Geometric Treatment." International J. of Transport Economics (forthcoming, 1982).
- Waters, W. G., II, and Woodland, A. D. "Econometric Analysis and Railway Costing." Monograph forthcoming: Centre for Transportation Studies, Vancouver, (1982).
- Wheaton, W. C. "Price Induced Distortion in Urban Highway Investment." Bell J. of Econ. (Autumn 1978), 622-32.
- Winston, C. "The Welfare Effects of ICC Rate Regulation Revisited." Bell J. of Econ. (Spring 1981).
- Zajac, E. E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." Amer. Econ. Review (Mar. 1970), 117-125.
- Zajac, E. E. Fairness or Efficiency: An Introduction to Public Utility Pricing. Ballinger, Cambridge (1978).

NETWORK MODELS IN TRANSPORTATION PLANNING
AND
THEIR INTERACTIVE/GRAPHIC IMPLEMENTATION

by

Michael Florian

Centre de recherche sur
les transports
Université de Montréal

NETWORK MODELS IN TRANSPORTATION PLANNING AND THEIR INTERACTIVE/GRAPHIC IMPLEMENTATION*

The interest in the mathematical modeling of transportation systems stems from the need to predict how people, vehicles and goods might make use of new or improved transport infrastructure in order to evaluate the benefits of the required investments. To this end it is necessary to build models of the demand for transportation and models that determine the way in which people, vehicles and goods use the transportation network. If such models may be constructed and their validity reasonably assured, then the prediction of the traffic flows on present and future transportation links may be carried out by simulating future situations and then evaluating the potential benefits of alternative improvement projects.

In the attempts that were made to construct mathematical models of route choice in transportation networks, the notion of equilibrium plays a central role. Suppose that the demand is known between an origin and a destination. Then it is of central importance to determine how these trips are attracted to the alternative routes available between origins and destinations. The idea of traffic equilibrium goes back to at least 1924, when Knight [22] gave a simple and intuitively clear description of a postulate of traffic behavior under conditions of congestion; to quote:

"Suppose that between two points there are two highways, one of which is broad enough to accommodate without crowding all the traffic which may care to use it, but is poorly graded and surfaced, while the other is a much better road, but narrow and quite limited in capacity. If a large number of trucks operate between the two termini and are free to choose either of the two routes, they will tend to distribute themselves

* Paper accompanied a seminar given to the Canadian Transport Commission on May 21, 1982. The interactive graphic system EMME/2 has been developed with the collaboration of H. Spiess, L. James-Lefebvre and A. Babin.

between the roads in such proportions that the cost per unit of transportation, or effective returns per unit of investment, will be the same for every truck on both routes. As more trucks use the narrower and better road, congestion develops, until a certain point it becomes equally profitable to use the broader but poorer highway."

Some 28 years later, Wadrop [34] stated the two principles that formalized this notion of equilibrium and introduced the alternative behavior postulate of the minimization of total costs. His first principle states that "the journey times on all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route". The implication of this principle is that the routes actually used are the shortest under the prevailing traffic conditions and their perception by the user. This characterization of equilibrium of route choice, or traffic equilibrium, which is identical to the notion described by Knight, became accepted over the past 30 years as a sound behavioral principle to describe the subdivision of trips over alternative routes. The traffic flows that satisfy this principle are usually referred to as "user optimized" flows, since each user chooses the route that he perceives to be the best. By contrast, the "system optimized" flows are characterized by Wardrop's second principle which states that "the average journey time is minimum". Wardrop compared these two principles and commented that the first criterion is quite a likely one in practice, since it implies an equilibrium situation in which no driver can reduce his journey time by choosing an alternate route, the second criterion is the most efficient since it minimizes the total travel time.

Although Wadrop discussed the equilibrium conditions for a general network, he did not propose a method to compute the corresponding flows. Beckmann, McGuire and Winsten [2] were the first to formulate and analyze a

Note: Numbers in square brackets appearing in the text refer to references.

Numbers in round brackets refer to the mathematical equations.

mathematical model of traffic equilibrium on a network. Their seminal work, which appeared in 1956, was the starting point for the contributions to follow in this area which were achieved mostly within the past 15 years.

The purpose of this seminar is to present the elements of the network models used in transportation planning, review their structural properties and most commonly used solution methods and outline current and potential applications.

THE FORMULATION OF NETWORK MODELS

The network models that are most commonly used are steady state models, in spite of the fact that almost all traffic phenomena are temporal. It is common to consider a given period of time, determine the demand characteristics during that period and determine the flow pattern that results from the interaction of the demand and the congestion characteristics of the transport infrastructure available.

Consider a transportation network which permits the flow of one type of traffic (vehicles or passengers) on its links. The nodes n , $n \in N$, represent origins, destinations and intersections of links, the links a , $a \in A$, represent the transportation infrastructure. The flow of trips on link a is v_a and the cost of travelling on a link is given by a user cost function $s_a(v)$, where v is the vector of link flows over the entire network. This cost function may model the time delay for travel on that arc, in which case it is commonly referred to as a volume/delay function. However, it may model other costs, such as fuel consumption. The vector user cost functions $s(v)$ is assumed to be monotone

$$(s(v') - s(v''))^T (v' - v'') \geq 0, \quad a \in A \quad (1)$$

continuous and differentiable.

The origin to destination demands, g_i , $i \in I$, where I is the set of origin/destination (O/D) pairs, may use directed paths k , $k \in K_i$, where K_i

is the set of paths for O/D pair i ($K_i \neq \emptyset$). The flows on paths k , h_k , satisfy conservation of flow and nonnegativity

$$\left. \begin{aligned} \sum_{k \in K_i} h_k &= g_i, \quad i \in I \\ \text{and } h_k &\geq 0, \quad k \in K_i, \quad i \in I \end{aligned} \right\} \quad (2)$$

The corresponding link flows v_a are given by

$$v_a = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} \cdot h_k, \quad a \in A \quad (3)$$

where $\delta_{ak} = \begin{cases} 1 & \text{if link } a \text{ belongs to path } k \\ 0 & \text{otherwise} \end{cases}$

The cost of each path $s_k (= s_k(v))$ is the sum of the user costs of the links in the path

$$s_k = \sum_a \delta_{ak} \cdot s_a(v), \quad k \in K_i, \quad i \in I \quad (4)$$

Let $u_i (= u_i(v))$ be the cost of the least cost path for any O/D pair i

$$u_i = \min_{k \in K_i} s_k, \quad i \in I \quad (5)$$

The demands g_i are given by functions $G_i(u)$, where u is the vector of least cost travel times for all the O/D pairs of the network

$$g_i = G_i(u), \quad i \in I \quad (6)$$

The vector demand functions $G(u)$ is assumed to be strictly antitone

$$(G(u') - G(u''))^T (u'_1 - u''_1) < 0, \quad 1 \in I \quad (7)$$

continuous, differentiable and bounded from above.

The network equilibrium model is formulated by supposing that for every O/D pair, Wardrop's user optimal principle is satisfied, that is

$$s_k^* - u_i^* \begin{cases} = 0 & \text{if } h_k^* > 0 \\ \geq 0 & \text{if } h_k^* = 0 \end{cases} \quad k \in K_1, \quad 1 \in I \quad (8)$$

over the feasible set (2)-(3).

It is relatively straightforward to show that (8) may be restated in the "complementarity" form

$$u_i^* \leq s_k^* \text{ and } (s_k^* - u_i^*) h_k^* = 0, \quad k \in K_1, \quad 1 \in I \quad (9)$$

and that (8) and (5) are equivalent to

$$s_{k_1}^* \leq s_{k_2}^*, \quad \text{if } h_{k_1}^* > 0, \quad k_1, k_2 \in K_1, \quad 1 \in I \quad (10)$$

Thus, Wardrop's first principle may be stated mathematically in several forms. However, the most useful restatement of the network equilibrium model is in the form of a variational inequality (Smith [32], Dafermos [7,8]). This is accomplished by noting that (8) is equivalent to

$$(s_k^* - u_i^*) (h_k - h_k^*) \geq 0, \quad k \in K_1, \quad 1 \in I \quad (11)$$

where h_k is any feasible path flows. If $h_k^* > 0$ then $s_k^* = u_i^*$ since h_k may be smaller than h_k^* . If $h_k^* = 0$,

then (11) is satisfied when $s_k^* - u_1^* \geq 0$. By summing over $k \in K_1$ and $i \in I$

$$\sum_{i \in I} \sum_{k \in K_1} s_k^* (h_k - h_k^*) \geq \sum_{i \in I} u_i^* (g_i - g_i^*) \quad (12)$$

Using (3) and (4) and by changing the order of summation on the left hand side it follows that

$$\sum_{a \in A} s_a(v^*) (v_a - v_a^*) \geq \sum_{i \in I} u_i^* (g_i - g_i^*) \quad (13)$$

Since the vector demand function is antitone, it is invertible. Let $W(g)=u$ be the inverse of the demand function. By making the corresponding substitution for u , the network equilibrium model may be stated as

$$\sum_{a \in A} s_a(v^*) (v_a - v_a^*) - \sum_{i \in I} W_i(g^*) (g_i - g_i^*) \geq 0 \quad (14)$$

over the feasible set (2)-(3).

It may be verified that (14) implies (8) by constructing a flow pattern which differs from the equilibrium flows on only one path k' , $k' \in K_1$, for which $h_{k'} = h_{k'}^* + \delta$, $0 \leq |\delta| < h_{k'}^*$. Applying (14) yields $s_{k'}^* \delta - W_1(g^*) \delta = 0$ which is equivalent to (8).

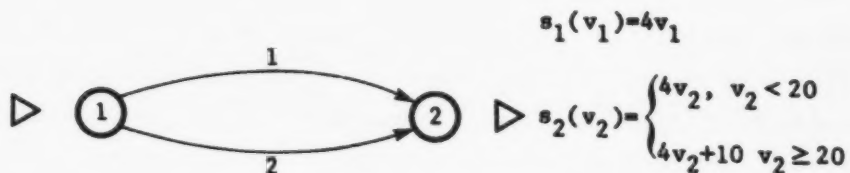
(14) may be rewritten by using matrix notation as

$$s(v^*)^T (v - v^*) - W(g^*)^T (g - g^*) \geq 0 \quad (15)$$

In the next section, a review of some mathematical properties of the network equilibrium model is presented.

SOME PROPERTIES OF THE NETWORK EQUILIBRIUM MODEL

The existence of a solution of the network equilibrium model is ensured by the continuity of the user cost and demand functions and by the fact that the feasible set (2)-(3) is compact if cycle flows do not occur and the demand functions are bounded from above. (See [1,21].) The following example illustrates that, when the user cost functions are not continuous, flows that satisfy the equilibrium conditions do not exist. The network consists of one O/D pair and two links as shown below.



The demand for 1 to 2 is 40. When $v_1 = v_2 = 20$, the cost of link 2 is higher than the cost of link 1, but for $v_1 = 20 + \epsilon$ and $v_2 = 20 - \epsilon$, $\epsilon > 0$, the cost of link 1 is higher than the cost of link 2.

In practice, the continuity requirement is not a hindrance, and in fact, all user cost functions that have been calibrated are continuous. The same is true of demand functions used in practice.

An important property of a network model that is used to simulate traffic flows for different future network configurations is that for a given scenario the flows and origin to destination costs be unique. Unless this were so, the comparison of the different future situations would be difficult to carry out; differences between scenarios would depend on which solution was found for each case as well as on the different network configurations. Fortunately, for many situations of interest, the network equilibrium model has unique flows, demand and origin to destination costs. This is ensured when the user cost functions are strictly monotone and the demand functions (and their inverses) are strictly antitone, as assumed in the model formulation.

To demonstrate this (see Smith [22], Dafermos [8], Aashtiani and Magnanti [1]), suppose that there exists two distinct solutions (v', g') and (v'', g'') . By writing (15) once with $v=v'$, $g=g'$ and $v^*=v''$, $g^*=g''$ and once with $v=v''$, $g=g''$ and $v^*=v'$, $g^*=g'$, one obtains by adding the two inequalities

$$(s(v') - s(v''))^T (v' - v'') - (W(g') - W(g''))^T (g' - g'') \leq 0 \quad (16)$$

However by (1) and (7) it follows that (16) is satisfied if and only if each term is equal to zero. Thus

$$(s(v') - s(v''))^T (v' - v'') = 0, \quad a \in A \quad (17)$$

$$\text{and} \quad (W(g') - W(g''))^T (g' - g'') = 0, \quad i \in I \quad (18)$$

It follows then from (18) that $g' = g''$, and from (17) that $s_a(v') = s_a(v'')$, hence the link user costs are unique as well. Since u_i^* , $i \in I$ are the lengths of the shortest paths computed for each O/D pair i based on the user costs $s_a(v^*)$, it follows that they are unique as well. One notes that if the user cost functions are strictly monotone, then the link flows are unique as well.

It is worthwhile to note that the path flows h_k are not unique, in general. Given the equilibrium link flows, the corresponding path flows are given as the solution of the simultaneous linear equations

$$v_a^* = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} \cdot h_k^*, \quad a \in A \quad (19)$$

Usually, the number of paths used is much larger than the number of links of the network, and in (19) the number of variables h_k^* far exceeds the number of constraints, which is equal to $|A|$. The consequence of this fact is that, in the analysis of particular network configurations, the paths used may be obtained, however, they do not represent reliable information, since they are not necessarily unique.

Another important property of the network equilibrium model is that it is stable. The equilibrium traffic flows depend continuously upon the travel demands and travel cost functions. Roughly speaking, small changes in the travel demands or in the user cost functions result in small changes in the traffic flows. This was demonstrated recently by Dafermos and Nagurney [6] for the general equilibrium model presented here and by Hall [20] for a version of the network equilibrium model that is discussed in the next section. This property is very desirable in its applications, provided that the model is a suitable representation of the observed route choice behavior.

SOLVING FOR THE EQUILIBRIUM FLOWS

Most of the applications of the network equilibrium model in practice have been achieved for simpler versions of the model (15) subject to (2)-(3). The user cost functions that have been calibrated are functions of only the flow in the link, that is $s_a(v) = s_a(v_a)$, $a \in A$ and the demand functions depend only on the travel cost between the origin and the destination of the node pair i , that is $G_i(u) = G_i(u_i)$, $i \in I$. The assumptions that were made on the function properties may be restated simpler to be a requirement that the user cost functions are all continuous and increasing and that the demand functions are continuous, strictly decreasing and bounded from above. The variational inequality (14) may now be restated as

$$\sum_{a \in A} s_a(v_a^*)(v_a - v_a^*) - \sum_{i \in I} W_i(g_i^*)(g_i - g_i^*) \geq 0 \quad (20)$$

over the set (2)-(3).

The application of this model in practice was facilitated by the equivalence between this variational inequality and a convex cost optimization problem for which the optimality condition is (20). This optimization problem is

$$\text{Min } Z = \sum_{a \in A} \int_0^v s_a(x) dx - \sum_{i \in I} \int_0^{g_i} w_i(y) dy \quad (21)$$

subject to (2)-(3).

Since the user cost functions are increasing and the demand functions are strictly decreasing, the objective (21) is convex and (20) is interpreted as the first order optimality condition, that the directional derivative of Z is larger than or equal to zero (see Property 6 of Appendix I). A rigorous proof of this equivalence may be found in Kinderlehrer and Stampacchia. [21] This is the classical result found by Beckmann et al. [2] The Kuhn-Tucker conditions of the convex cost optimization problem (21) subject to (2)-(3) (see Appendix III) are equivalent to the statement of Wardrop's user optimal principle of route choice in the form (8).

The existence of the equivalent optimization problem greatly facilitated the implementation of the network equilibrium model in practice, since a number of efficient algorithms are available for solving such problems, while the solution of variational inequalities of large scale is more difficult. It may be said without reservation that most of the practical applications of the network equilibrium model are based on the initial work of Beckmann et al. even though recently, the variational inequality formulation opened the way to understanding and solving network equilibrium models of increasing complexity.

From the large repertory of methods that are available for solving convex cost minimization problems subject to linear constraints, the adaptation of the linear approximation method of Frank and Wolfe [18] and

some of its variants is the most commonly used algorithm. Bruynooghe, Gibert and Sakarovitch [4] were the first to propose the method, however the subsequent work of LeBlanc, Morlok and Pierskalla [24] and Nguyen [29] made this method popular in practice. (See also [25,30] and [9] for a different approach.) Computer codes are widely available for solving the fixed demand version of this problem, that is when $g_i = \bar{g}_i$, $i \in I$. We proceed to present the details of the adaptation of this method to solve the fixed demand network equilibrium problem and then discuss some of its variants, for both the fixed and variable demand versions.

The fixed demand network equilibrium problem will now be considered:

$$\text{Min } Z = \sum_{a \in A} \int_0^{v_a} s_a(x) dx \quad (22)$$

$$\text{subject to} \quad \sum_{k \in K_i} h_k = \bar{g}_i, \quad i \in I \quad (23)$$

$$h_k \geq 0, \quad k \in K_i, \quad i \in I \quad (24)$$

$$\text{and one recalls (3), that is } v_a = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} \cdot h_k, \quad a \in A.$$

As described in Appendix II, starting from an initial feasible solution the linear approximation method obtains a feasible direction by linearizing the objective function, solving a linear programming subproblem and then finding an improved solution on the line segment between the current solution and the solution of the subproblem. For the network equilibrium problem the linearized subproblem at iteration ℓ is:

$$\text{Min } \sum_{a \in A} s_a(v_a^\ell) \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} \cdot f_k \quad (25)$$

$$\text{subject to } \sum_{k \in K_i} f_k = \bar{g}_i, \quad i \in I \quad (26)$$

$$f_k \geq 0, \quad k \in K_i, \quad i \in I \quad (27)$$

By changing the order of summation in (25) and by using (4), the objective becomes

$$\text{Min } \sum_{i \in I} \sum_{k \in K_i} s_k^\ell \cdot f_k \quad (28)$$

The solution of this problem is obtained trivially by computing the shortest paths for each O/D pair i and allocating the demand $\bar{g}_i$ to that path. Note that the complementarity slackness conditions for the dual constraints of this linear program are

$$f_k (u_i - s_k^\ell) = 0, \quad k \in K_i, \quad i \in I \quad (29)$$

and the dual constraints are

$$u_i \leq s_k^\ell, \quad k \in K_i, \quad i \in I \quad (30)$$

The optimal u_i are the length of the shortest paths based on the current link user costs $s_a (v_a^\ell)$ and (29) is a statement of Wardrop's user optimal principal when the link costs are constant.

The solution to the problem (25)-(27) is then given by

$$f_a^\ell = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} f_k \quad (31)$$

and the direction of descent is simply $d_a^l = (f_a^l - v_a^l)$.

The shortest path calculations required in the solution of this subproblem may be carried out by any of the efficient methods currently available.

An iteration of the linear approximation algorithm is completed by finding the solution of

$$\text{Min}_{0 \leq \lambda \leq 1} \sum_a \int_0^{v_a^l + \lambda d_a^l} s_a(x) dx \quad (32)$$

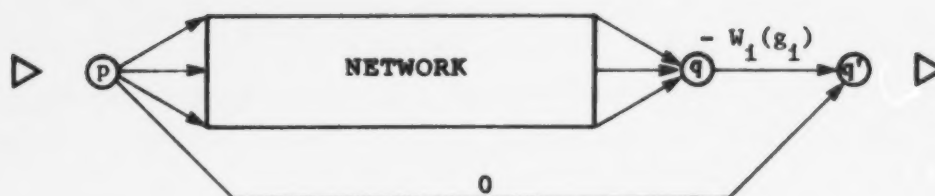
or equivalently annulling its gradient, that is find λ^l for which

$$\sum_{a \in A} s_a(v_a^l + \lambda^l d_a^l) d_a^l = 0 \quad (33)$$

For numerical reasons it is preferable to find λ by (33). The linearization is repeated for $l = 1, 2, \dots$ until a satisfactory solution is obtained.

This algorithm has several marked advantages. The necessary paths are generated as required and need not be kept from one iteration to the next. Thus, the storage requirements are modest and do not increase with the number of iterations. At each iteration only the arc flows v_a^l need to be kept, in addition to the link costs $s_a(v_a^l)$ and the network. A drawback of this algorithm is that it converges slowly in the vicinity of the optimal solution. In practice this is not very limiting, since very accurate solutions are not needed for large scale problems.

The variable demand problem is just slightly more difficult to solve than the fixed demand problem. From a pedagogical viewpoint, the following transformation of the variable demand problem into a fixed demand problem is instructive (see Dantzig *et al.* [10]). For each O/D pair i , a node and two arcs are added as indicated below



For the pair i , the origin node is p and the destination node is q . A node q' and the arcs (p, q') and (q, q') are added. The cost of arc (p, q') is zero and the user cost function on arc (q, q') is $-W_1(g_1)$. A demand of $g_1^{\max}$, which is an upper bound on the equilibrium demand is imposed from p to q . The solution of this fixed demand problem has the property that all used paths are of length 0; thus any path on the original network that carries flow has length u_1^* and therefore $u_1^* = W_1(g_1^*)$ or $g_1^* = G_1(u_1^*)$.

This transformation is not efficient computationally since the arc (p, q') is not necessary in the straightforward adaptation of the linear approximation method to solve the variable demand problem, which only requires the addition of the arc (q, q') , or for all O/D pairs, the addition of a matrix of dimensions $|I| \times |I|$ which contains the current O/D matrix at each iteration. For the variable demand network equilibrium problem the linearized problem at iteration ℓ is

$$\text{Min} \sum_{a \in A} s_a(v_a^\ell) \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} \cdot f_k - \sum_{i \in I} W_1(g_1^\ell) w_i \quad (34)$$

$$\text{subject to} \quad \sum_{k \in K_i} f_k - w_i = 0, \quad i \in I \quad (35)$$

$$f_k \geq 0, \quad k \in K_i, \quad i \in I, \quad w_i \geq 0, \quad i \in I \quad (36)$$

As before, the objective (34) simplifies after some manipulations to

$$\text{Min} \sum_{i \in I} \left(\sum_{k \in K_i} s_k^\ell f_k - W_1(g_1^\ell) w_i \right) \quad (37)$$

which may be solved by determining u_i to be lengths of the shortest paths based on the current link cost $s_a(v_a^l)$ and determining the demand to be allocated to these shortest paths as follows, where $g_i^{\max}$ is an upper bound of g_i .

$$w_i^l = \begin{cases} g_i^{\max} & \text{if } u_i < w_i(g_i^l) \\ g_i^l & \text{if } |u_i - w_i(g_i^l)| < \epsilon \\ 0 & \text{if } u_i > w_i(g_i^l) \end{cases} \quad i \in I \quad (38)$$

The descent direction is $\{(f_a^l - v_a^l), (w_i^l - g_i^l)\}$ where f_a^l is obtained by assigning the demand w_i^l to the shortest paths. The step size is determined by finding λ^l which annuls the gradient

$$\sum_{a \in A} s_a(v_a^l + \lambda(f_a^l - v_a^l))(f_a^l - v_a^l) - \sum_{i \in I} w_i(g_i^l + \lambda(w_i^l - g_i^l))(w_i^l - g_i^l) \quad (39)$$

An interesting variant of the linear approximation method for solving the variable demand network equilibrium problem is that of partial linear approximation that was first suggested by S. Evans [14] for solving the combined trip distribution/assignment problem. This problem is a variant of the variable demand problem where the demand functions are taken to be a negative exponential function of the origin to destination travel cost, that is $G_i(u_i) \approx \exp(-\theta u_i)$ and additional constraints are imposed on the total trips originating at origins p and terminating at destinations q . This model may be stated as

$$\text{Min } \sum_{a \in A} \int_0^v s_a(x) dx + \theta \sum_{i \in I} g_i \ln g_i \quad (40)$$

$$\text{subject to } \sum_{k \in K_i} h_k - g_i = 0, \quad i \in I \quad (41)$$

$$\left. \begin{aligned} \sum_p g_{pq} &= D_q, \quad \forall q \\ \sum_q g_{pq} &= 0, \quad \forall p \end{aligned} \right\} i=(p,q) \quad (42)$$

$$h_k \geq 0, \quad k \in K_1, \quad i \in I; \quad g_i \geq 0, \quad i \in I \quad (43)$$

See also [17]. Although well studied, this model has not yet found application in practice.

Evans' algorithm was interpreted by Stewart [33] to be a partial linear approximation, where, for the variable demand problem the user cost functions are linearized, but the inverse demand functions are not. The resulting subproblem is

$$\text{Min} \sum_{a \in A} s_a(v_a^l) \sum_{i \in I} \sum_{k \in K_1} \delta_{ak} \cdot f_k - \sum_{i \in I} w_i(w_i) \quad (44)$$

$$\text{subject to} \quad \sum_{k \in K_1} f_k - w_i = 0, \quad i \in I \quad (45)$$

$$f_k \geq 0, \quad k \in K_1, \quad i \in I; \quad w_i \geq 0, \quad i \in I \quad (46)$$

This subproblem is still easy to solve by determining u_i to be the shortest paths based on the current link cost $s_a(v_a^l)$ and then determining the demand to be allocated to these shortest paths as follows

$$v_i^l = C_i(u_i), \quad i \in I \quad (47)$$

Intuitively, this direction should be better, since the subproblem used for finding the direction is "closer" to the original problem. Computational results [19,23] support this intuition. As this subproblem is not more difficult to solve than the linear approximation subproblems, it is attractive to use in practice.

Another variant of the linear approximation method which is particularly useful for the variable demand version of this problem is the parallel tangent or PARTAN variant. At each step of the linear approximation algorithm, another direction of descent is generated by every other solution v^{l-2} and v^l . Thus the new solution is generated by finding $\alpha^l, \alpha^l \leq \bar{\alpha}$ which minimizes the objective function for the solution $v^{l-2} + \alpha(v^l - v^{l-2})$, where $\bar{\alpha}$ is the largest step size allowed in order to maintain the nonnegativity of the path flows. Note that this direction is not guaranteed to be a descent direction and, in particular if $\alpha^l < 1$, the objective function decreases in the opposite direction. The implementation of this variant requires the storage of all the weights that are given to each solution of the direction finding subproblem, of the step sizes α^l taken in the PARTAN direction and of the solutions v^{l-1} and v^{l-2} . The computational results obtained with this method suggest that it is a worthwhile modification of the basic method. Its additional storage requirements are modest for the fixed demand network equilibrium problem and for problems of moderate size, it may be suitable for the variable demand problem as well.

SYSTEM OPTIMAL VS USER OPTIMAL FLOWS

Wardrop's second principle is usually referred to as yielding "system optimal" flows since these are the flows which minimize the total cost in the system. The fixed demand version of this problem is

$$\text{Min } \sum_{a \in A} v_a \cdot s_a(v_a) \quad (48)$$

subject to (23), (24).

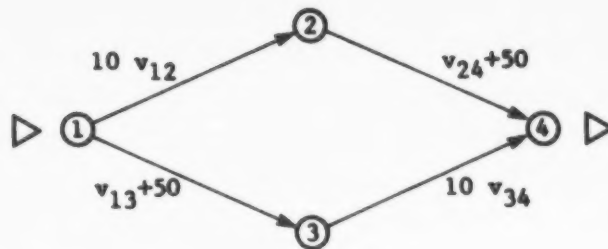
By applying the Kuhn-Tucker conditions to the system optimal problem, one obtains the system optimal flows that have equal marginal cost on all the used paths, that is

$$m_k = \sum_a \delta_{ak} (s_a(v_a) + v_a \cdot s'_a(v_a)) \quad (49)$$

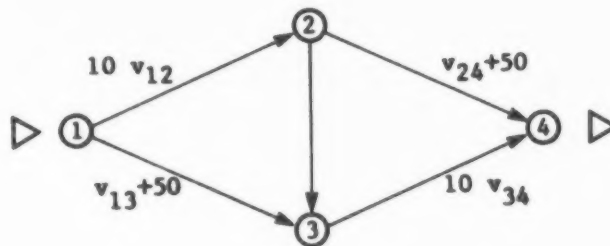
$$\text{and } m_k^* - u_1^* \begin{cases} = 0 & \text{if } h_k^* > 0 \\ \geq 0 & \text{if } h_k^* = 0, k \in K_1, i \in I \end{cases} \quad (50)$$

The computation of the system optimal flows may be carried out by the linear approximation algorithm by making the appropriate modification of objective function and use of marginal link costs in the direction finding subproblem. Since the system optimal flows are more "efficient", in the sense that the trips are undertaken with a minimum total cost, than the user optimal flows, if link tolls equal to $v_a \cdot s'_a(v_a)$ could be imposed, then the users seeking their best routes will distribute themselves over the paths of the network in a system optimal way. The practical difficulty with imposing such tolls is that any tolls are likely to change the original user cost functions as well. Economists refer to such tolls as marginal cost pricing. The notion of marginal cost pricing may be found as early as 1844 with the work of Dupuit. [13] See also [27].

In general, the user optimal flows are different than the system optimal flows in a given network. For large, uncongested networks the difference may be small, since for constant link travel costs there is no difference between the user and system optimal flows. However, these differences are significant when evaluating different scenarios where link additions or deletions are considered. Braess [3] (see also Murchland [26]) provided a classical example of such differences, by using the network shown below where the user cost functions are indicated on the links.



The equilibrium flows are three units on the path (1,2)(2,4) and three units on the path (1,3)(3,4). The total cost is $6 \times 83 = 498$. If the link (2,3) is now added to the network with user cost function $v_{23} + 10$



the equilibrium flows are two units on each of the paths of the network with total cost of $6 \times 92 = 552$. Thus, the addition of an arc increases the total travel cost as well as the travel cost to each user. Similarly, the suppression of the arc (2,3) would reduce the total travel cost and the cost to each user. This apparent paradox is understood by examining the system optimal solution which is three units each along the paths (1,2)(2,4) and (1,3)(3,4), even when the arc (2,3) is present. There is no reason to anticipate that total travel cost may be reduced when a link is added, since these costs are not minimized by the user optimal flows.

The occurrence of the Braess' paradox is of significant importance in the decisions taken regarding the additions or deletions of links in a network. If the user optimal principle is suitable as the behavioral representation of the route choice, then given a particular network, it may be advantageous to restrict the use of some of its links and care must be exercised to properly evaluate link additions to detect possible occurrences

of increases in total costs and increase of service levels for some of the users. The Braess' paradox may certainly occur in real networks, as evidenced by the work of Fisk and Pallottino. [15]

The other relevant question related to system optimal flows is the form of the problem with variable demand. One appropriate objective function of this model is

$$\text{Min } \sum_{a \in A} v_a \cdot s_a(v_a) - \sum_{i \in I} \int_0^{g_i} w_i(y) dy \quad (51)$$

subject to (2), (3).

By applying the Kuhn-Tucker conditions to this problem one finds (50) and

$$\left. \begin{array}{l} G_i(u_i^*) = g_i^* \\ \text{where } u_i^* = \min_{k \in K_i} m_k^* \end{array} \right\} i \in I \quad (52)$$

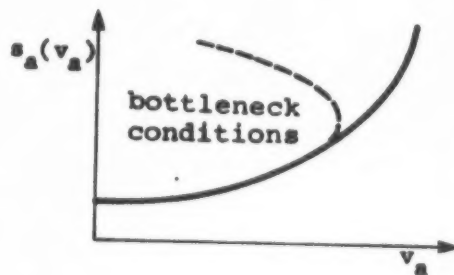
that is, the cost perceived by the user is the marginal cost, rather than the average cost. This model is suitable for applications in rail networks where it may be argued that the tariffs are proportional to the marginal costs.

APPLICATION AND VALIDATION OF THE NETWORK EQUILIBRIUM MODEL

The validation of the network equilibrium model is based at present on relatively scarce empirical studies. Florian and Nguyen [16] applied the model to the urban network of the City of Winnipeg and Dow and Van Vliet [12] applied it to the City of Leeds. A practical problem in applying the network equilibrium model is the determination of the origin/destination matrix. Even when survey data is available, it does not include information for all the trips taken during the peak hour and the flow predictions must be reconciled with observed link counts. The process of calibrating a network equilibrium model involves the network representation, the calibration and assignment of user cost functions to the links of the network and possible

adjustments of the origin/destination matrix. It may be stated that some significant approximations are often made in order to build the necessary data base for the application of the model. Yet, it may be the best predictive tool available today for predicting the impact of network changes in the short and medium term. The link flows obtained in the studies [16,12] simulate flows on congested links quite well and the origin to destination travel times are satisfactorily reproduced as well.

The application of the network equilibrium model is based on the assumption that the traffic is under free flow conditions and that no severe bottlenecks occur. When this assumption is not satisfied, increasing user cost functions do not model properly the phenomenon of increased travel time and reduced flow on links subject to bottlenecks, as indicated in the graph below.



The use of the network equilibrium model is not particularly expensive in computer expenditures. Experience with network of medium size (200 origins/destinations, 1 500 nodes and 4 000 links) indicate that, on a typical mainframe computer, the cost of 10-15 iterations of the linear approximation method is in the range of \$150 - \$300. The possibility of carrying out these computations on the new generation of 16 bit micro-computers will probably render the cost even lower.

The network equilibrium model may also be used to measure fuel consumption impacts by incorporating the empirical fuel consumption equations of Chang et al. [5] in an appropriate way.

MULTIPATH FLOWS WITH CONSTANT LINK COSTS

The user equilibrium flows are suitable for representing route choice behavior when congestion effects significantly increase travel costs. When the link user costs are constant, the user optimal principle is satisfied by assigning all the flows to the shortest cost path between origins and destinations (recall (25)-(27) and (28), (29)). Since it is seldom in practice that travellers have exact knowledge of the link times, it is common that more than one route is used, even when no congestion is present. Dial [11] proposed a route choice model based on the logit model which postulates that

$$h_k = \bar{g}_i \cdot \exp(-\theta s_k) / \sum_{k \in K_i} \exp(-\theta s_k) \quad (53)$$

subject to the usual flow conservation and nonnegativity constraints (23), (24). This model has the equivalent optimization problem

$$\text{Min} \sum_{i \in I} \sum_{k \in K_i} h_k (\ln h_k + \theta s_k) \quad (54)$$

subject to (23), (24). The solution of this model requires the explicit storage of the paths. Dial ingeniously circumvented this problem by restricting the paths to be considered and succeeded in solving this restricted version of the problem without requiring path enumeration.

Although this model found much use in practice, the use of the logit path choice makes the resulting flows very sensitive to the network representation and the restriction of the paths results in the model being unstable; slight modifications in the network may induce large changes in the resulting flows.

APPLICATION VIA INTERACTIVE GRAPHIC NETWORKS

The network models that are used in transportation planning have undergone rapid development in the past decade. A wide variety of transportation planning methods have become available for various urban and regional approaches. Most of these methods have been implemented for use on modern computers in a "batch" type environment. Although many successful applications of these methods have been achieved, and the new generation of urban transportation models (such as UTPS, EMME/1) go a long way to alleviate the difficulties associated with their behavioral validity and computational efficiency, the dialogue with the planner is still rather cumbersome. The transfer of information between person and computer is done via relatively cumbersome procedures, sometimes involving computer analysts who have no interest in planning, and are also limited by the management of a given computer installation. The typical situation that requires the use of such computer based models is the evaluation of future scenarios, which may reflect changes in the road and transit networks or changes in the socio-economic characteristics of the urban area. Such use requires quick and efficient communication with the computer based models which "evaluate" the proposed scenarios.

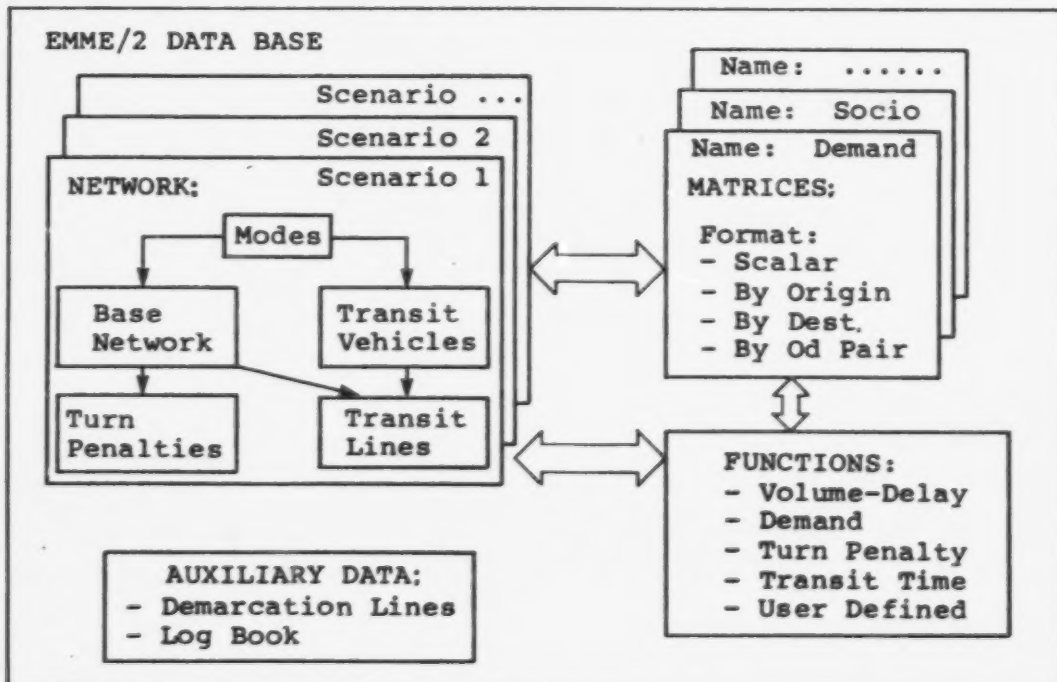
The approach that has emerged over the past few years, particularly in the context of transit route planning, is to use interactive graphics to enable direct dialog, with real time graphic or non-graphic response, between the planner and the computer based planning method. It allows the planner to engage in the planning process by using the terminology that he is familiar with and obviates data processing type tasks. Once the data base is set up, the planner need not know or concern himself with the technical details of computer programming or computer systems. It also permits him the instantaneous visualization of input data, results of computations and information retrieved from the data base, all in public or list form. EMME/2 is a multi-mode urban transportation planning method designed for interactive use. It is more comprehensive than other interactive graphic methods that have been developed to date.

GENERAL CONCEPTS THAT UNDERLY THE DESIGN OF EMME/2

The data that may be entered, modified and used for calculations in EMME/2 falls into three main categories: networks, matrices and functions. Diagram 1 gives a schematic representation of the EMME/2 data base, and is helpful in clarifying the descriptions that follow.

The network data may be input, modified and displayed by using the Network Editor of EMME/2. The matrices of the EMME/2 data base may be entered, modified, displayed and various matrix calculations may be performed by using the Matrix Editor. All the relevant functions of the EMME/2 data base are entered and evaluated by using the Function Editor. The EMME/2 data base contains Demarcation Lines and a Log Book, which is referred to as the auxiliary data. Demarcation lines may be super-imposed on graphic displays to make them more readable and a Log Book keeps track of the use made of the various EMME/2 modules in manipulating the data base. Brief descriptions of the networks, the matrices and the functions that may be entered and manipulated by the EMME/2 system follow.

Diagram 1



The Networks

The transportation infrastructure that spans the region studied is represented by a multimodal network. The network descriptors are the modes, base network, transit vehicles, transit lines and turn penalties. Any of this data may be modified, at any time, provided that the hierarchy of the data structure is represented, (e.g., a transit line cannot use a vehicle type not present in the vehicle table). See Diagram 2.

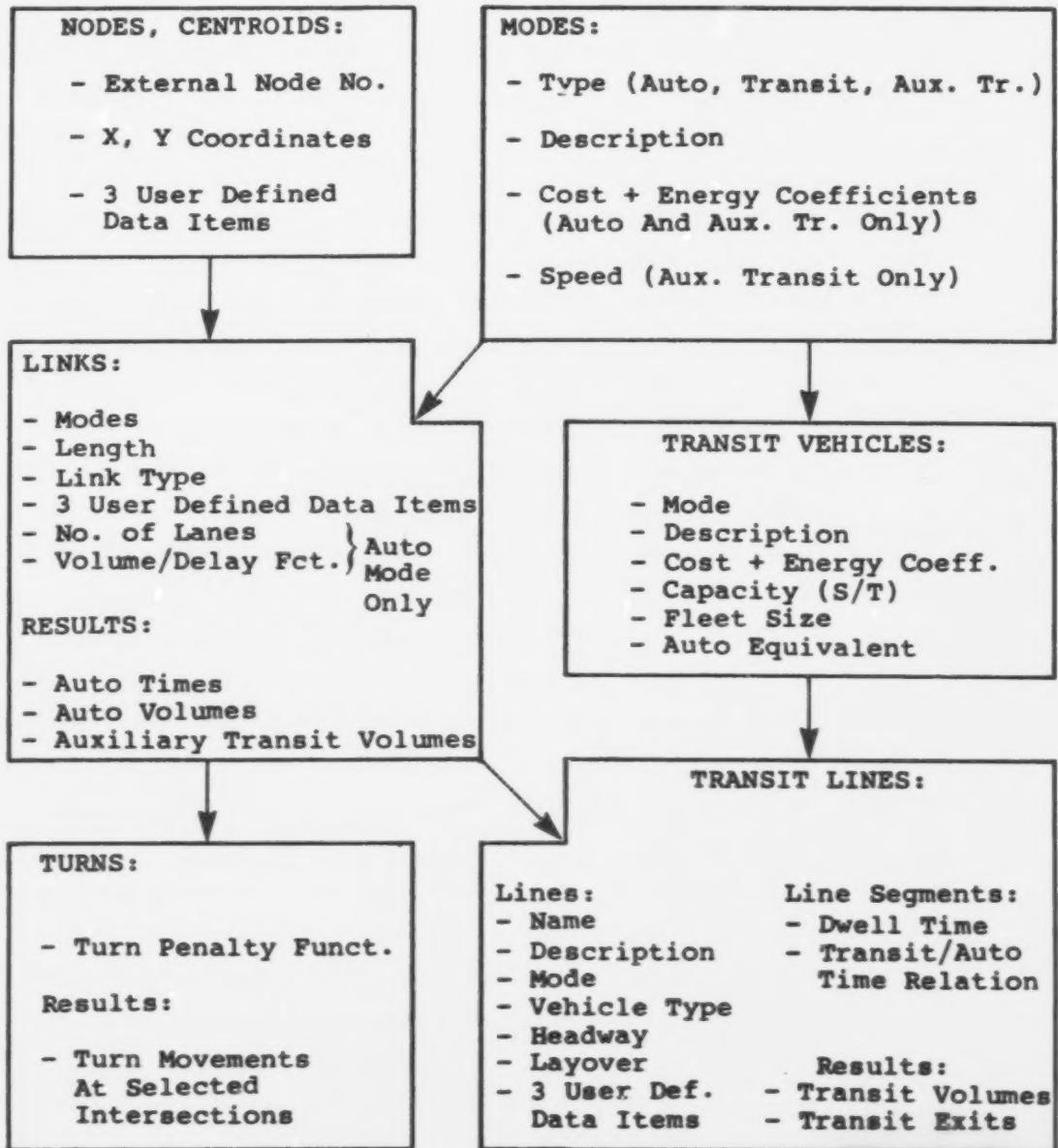
The base network is defined by the list of all nodes and all the links between those nodes, used by any of the modes. Intersections that are described at the level of detail of turning movements are prepared separately. The "turn penalties" are indicated only for penalized movements.

The line itinerary is defined as a sequence of nodes. It is not necessary to specify all the nodes along a route. If nodes are omitted, the line is assumed to pass through the nodes on the shortest path between the specified nodes. Dwell time and the transit time function, need only be specified when their value changes. This data is stored by line segment. Up to two layovers are allowed in a line itinerary and they are specified immediately after the node where the layover takes place. It is not necessary for a transit line to return to its starting point, such as certain services during the peak hour. The base network is defined by the unit of all nodes and all the links between those nodes, used by any of the modes.

The user may specify up to three additional data elements for each node, link and transit line. Such "user data" may include observed link flows, observed link travel times or any other data that the planner intends to use in subsequent analyses. Another example would be accident rates at given nodes of the network.

Each complete network data set (modes, base network, transit vehicles, turn penalties, transit lines) makes up a scenario, and the base year is one of these scenarios. The user may define a new scenario, by duplicating the data present in an existing scenario and making the appropriate changes in any one of the network data components. All these

Diagram 2
EMME/2 NETWORK STRUCTURE:



manipulations may be performed interactively, with graphical output, when appropriate. In EMME/2, the network data that corresponds to a "base year" is not conceptually different from that corresponding to a future year configuration.

The Matrices

The matrices that are handled in EMME/2 may be full matrices, origin or destination vectors or scalars. These may contain various socio-economic data related to the zone subdivision of the urban area studied, such as origin/destination demand matrices, origin/destination travel times by mode, etc. A matrix may be both an input to or an output from a computational procedure.

The matrix editor has been designed in a way that permits the user to choose a matrix format appropriate to his data, in order to avoid duplication of data in the data base. For instance, if a full matrix has the same value in all its cells, then the user may specify a scalar matrix, give the value and the matrix editor will use only the space required to keep the scalar in the data base but when a full matrix is requested in an EMME/2 module, the matrix editor will automatically expand the scalar value to a full matrix with that value in all of its cells within the module at execution time. The same concept applies to given values for origins or destinations.

The matrix editor does not define any matrices by default. The definition of a matrix and its contents are left to the user. Thus, if an assignment routine expects the availability of an origin/destination matrix, the user defines this matrix and ensures that it contains the correct data. This provides great flexibility in evaluating a given scenario with different O/D matrices.

Any network scenario created by the network editor may be used with any of the matrices present in the data base. The user identifies the relevant matrices that he wishes to use with a given scenario. The matrix editor of EMME/2 also contains matrix balancing procedures in two and three

dimensions. Their judicious use, with the appropriate matrices permits a user to scale matrices or to implement some versions of classical spatial interaction models.

Whenever the matrix editor expects a zone number, a zone group can be specified, indicating that the operation applies to all zones within the group. A zone group is a set of zones. A set of zone groups that includes all the zones is called an ensemble. Within an ensemble, a zone may be in only one group. The user of the EMME/2 system may specify up to 26 ensembles, each containing up to 100 zone groups. The advantage of using zone groups is that for many applications it is necessary to select subsets of the zones and using zone groups, this selection may be accomplished efficiently.

The matrix editor includes a module which permits the user to perform matrix calculations. The user specifies the operations that he wants to perform as an algebraic expression. Such algebraic expressions may be stated by choosing from 16 logical and algebraic operands and nine intrinsic functions. The variables that may be used in these expressions are the matrices of the EMME/2 data bank, centroid numbers and zone groups.

The Functions

All functions that are used in EMME/2, such as volume/delay functions, transit/auto travel time relationships, turn penalties and demand functions are specified by the user as algebraic expressions. These functions are not part of the code, as "user defined" functions or subroutines, as in the EMME/1 system or other similar systems. When an EMME/2 module requires the use of a particular function, the function is then evaluated by using the appropriate data that corresponds to the variables specified by the user in the algebraic expressions that define these functions.

In addition to the classes of functions described above, the user is free to define and display functions that are not employed in any one of the standard calculations of the EMME/2 system.

Assignment Procedures

The most general assignment procedure provided by EMME/2 is a multimodal equilibrium assignment method with variable demand. It computes the equilibrium demands, flows and service levels for all the modes considered. As the assignment procedures have a modular structure, the user may select equilibrium assignment with fixed or variable demand on the road network, or variants of shortest path or multi-path assignment on the network served by the transit and any or all of the auxiliary transit modes. For variable demand assignment, both mode choice and direct demand functions may be specified.

The Results

EMME/2 permits the user to obtain a wide variety of results, both in interactive graphic form and/or as a printed output. The main feature of the results is interactive comparison of scenarios with accompanying graphical display. While the main results pertaining to comparison of scenarios are related to link flows, origin/destination demands and service levels, a wide variety of other results may be obtained using the user defined data (e.g., comparison of predicted vs. observed flows for calibration).

Worthy of emphasis is that unlike a "batch" code, where each successful execution terminates with a particular set of results, an interactive graphic code permits the user to obtain results of different types during a particular EMME/2 session. The notion of "result" is thus different from that of a batch code and may be considered to consist of the entire gamut of displays, results of computational procedures, data bank queries and scenario comparisons.

Demarcation Lines and Log Book

For a given urban area the user may define demarcation lines, that may be superimposed on a graphical output that covers the area spanned by the zone subdivisions. A demarcation line may identify geographical characteristics of the urban area, such as rivers or mountains or certain regions of the city, such as the central business district.

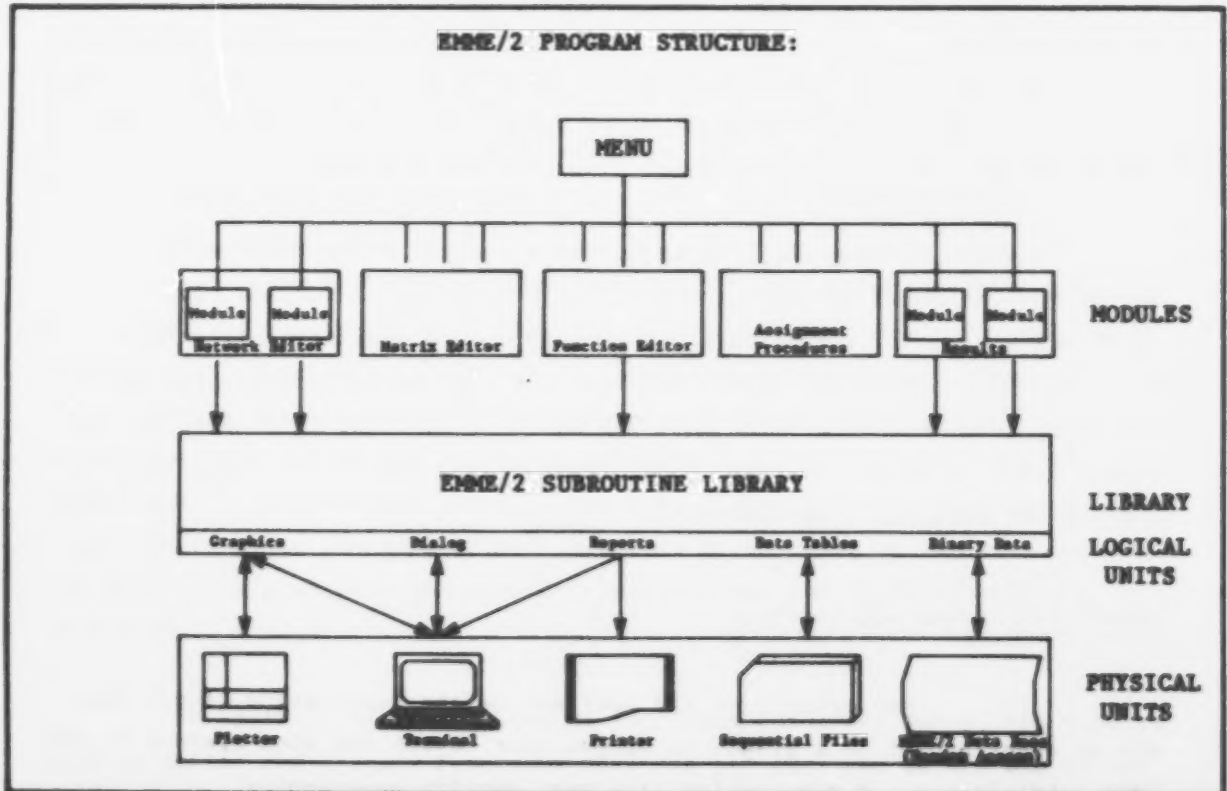
An automatic Log Book keeps record of the identity of the user and of the modules and elements of the data base user during all EMME/2 sessions.

EMME/2 - THE COMPUTER BASED SYSTEM

EMME/2 has been coded in standard ANSI Fortran IV and has been designed for each transferability to various computer makes. It has been developed on the Cyber 173 of the Université de Montréal, however it may be adapted easily for other computers, including a certain class of mini (micro) computers.

The EMME/2 code has a modular structure. Each of the modules is an independent program but all the modules share the EMME/2 subroutine library. All data transfers between modules occur only via the data base. Diagram 3 gives a schematic representation of the program structure.

Diagram 3



At present, EMME/2 is implemented for use on any of the Tektronix 4010 and 4110 series of terminals. Any output displayed on the screen may be copied by using Tektronix 4631 hard copy unit. The same output may be drawn with a high technical quality by using the Tektronix 4663 digital plotter, when available. EMME/2 may be easily adapted for use with other graphic equipment of comparable resolution.

SOME OUTPUT EXAMPLES OF EMME/2

The figures that are presented in this section of the report are a sample of the kind of output that is made possible by the EMME/2 system.

Figure 1: The menu which displays the available EMME/2 modules.

Figure 2: Plot of a multimodal base network. There are three modes C (car), B (bus) and A (access), the CBD is not displayed and two rivers serve as demarcation lines.

Figure 3: An example of the dialog used to plot a window of the base network.

Figure 4: The window of the base network that is plotted as a result of the dialogue of Figure 3. The plot displays node numbers and also the number of lanes and the volume/delay function index for each link.

Figure 5: Plot of subnetwork of the multimodal base network displayed in Figure 2 which corresponds to district 3. The district is identified by link types 31 to 39.

Figure 6: The graphic work sheet which is used by the Interactive Network Editor. The window of the base network which is about to be edited is displayed. The upper part of the graphic work sheet contains the command area where the user indicates the data tables to be modified and the operations that will be used in the modifications.

Figure 7: The graphic work sheet after certain modifications of the window of Figure 6 have been carried out. The left hand margin is a "pad" which records the modifications and permits an abbreviated dialogue for additions. For instance, the data of link (1 020, 1 018) was first listed, then links (1 020, 1 018) and (1 018, 1 016) were deleted, then node 2 000 was added, etc...

- Figure 8: Plot of shortest distance path between two nodes on the base network. The two nodes are identified interactively by using the cross hair cursors. The right hand margin contains the "from node" the "to node" and the distance in unit length. The shortest path is displayed with a heavy line on the base network.
- Figure 9: Plot of transit line number 3 on a window of the base network which displays only links used by the bus (B) mode.
- Figure 10: Plot of all the transit lines that service district number 3.
- Figure 11: Plot of turning movements at selected intersections. The forbidden turns are not shown and the penalized turns are indexed with the number of corresponding penalty function.
- Figure 12: Plot of productions (row totals) and attractions (column totals) of the origin/destination matrix. The right hand margin gives the scale used for the bars of the display.
- Figure 13: Plot of a column of the origin/destination matrix or "desire lines" for centroid four. The right hand margin gives the scale used for the bars of the display.
- Figure 14: Plot of four volume/delay functions. The right hand margin gives the functions plotted (FD1, FD2, ...) and the parameters that were fixed at certain values (VOLTR=0, LANES=2, LENGTH=.5).
- Figure 15: Plot of a shortest path tree on the auto network (including penalized turns), based on link times that result from an assignment. The right hand margin gives the root of the tree, centroid 14.
- Figure 16: A comparison of two histograms of travel time distributions. The "base year" corresponds to the travel time matrix TRTIMO and the demand matrix GPQTR0, while the "future scenario" corresponds to travel time matrix TRTIM1 and the demand matrix GPQTR1. The right hand margin contains as well the mean and standard deviation of each distribution.
- Figure 17: Plot of auto link volumes that result from an assignment on the auto network. The right hand margin indicates that volumes smaller than 100 are not plotted (THRESHOLD : LOWER : 100) and the scale used for the bar width on the display.

Figure 18: Plot of auto link speeds as a result of an assignment on the auto network.

Figure 19: Plot of transit volumes on transit line number 41 as a result of an assignment on the transit network.

Figure 1

1. UTILITIES

- 1.11 DIRECT DATA BASE MANIPULATIONS
- 1.21 LOG BOOK CONSULTATION
- 1.22 SCENARIO MANIPULATIONS
- 1.23 MODIFY TITLES AND UNITS
- 1.31 I/O FILE MANAGEMENT
- 1.41 INPUT / MODIFY / DISPLAY DEMARCATION LINES

2. NETWORK EDITOR

- 2.01 INPUT / MODIFY / OUTPUT MODES
- 2.02 INPUT / MODIFY / OUTPUT TRANSIT VEHICLES
- 2.11 INPUT BASE NETWORK USING BATCH ENTRY
- 2.12 INPUT / MODIFY BASE NETWORK INTERACTIVELY
- 2.13 PLOT BASE NETWORK
- 2.14 OUTPUT BASE NETWORK
- 2.15 PLOT SHORTEST PATHS ON BASE NETWORK
- 2.21 INPUT TRANSIT LINES USING BATCH ENTRY
- 2.22 INPUT / MODIFY TRANSIT LINES INTERACTIVELY
- 2.23 PLOT TRANSIT LINES
- 2.24 OUTPUT TRANSIT LINES
- 2.31 INPUT / MODIFY / DISPLAY TURN PENALTIES
- 2.41 INPUT / MODIFY / DISPLAY USER DEFINED DATA

3. MATRIX EDITOR

- 3.01 INPUT / MODIFY / DISPLAY ZONE GROUPS
- 3.11 INPUT MATRICES USING BATCH ENTRY
- 3.12 INPUT / MODIFY MATRICES INTERACTIVELY
- 3.13 PLOT MATRICES
- 3.14 OUTPUT MATRICES
- 3.15 COMPARE MATRICES
- 3.16 PLOT HISTOGRAM OF MATRICES
- 3.21 MATRIX CALCULATIONS AND BALANCING

4. FUNCTION EDITOR

- 4.11 INPUT FUNCTIONS USING BATCH ENTRY
- 4.12 INPUT / MODIFY FUNCTIONS INTERACTIVELY
- 4.13 PLOT FUNCTIONS
- 4.14 OUTPUT FUNCTIONS

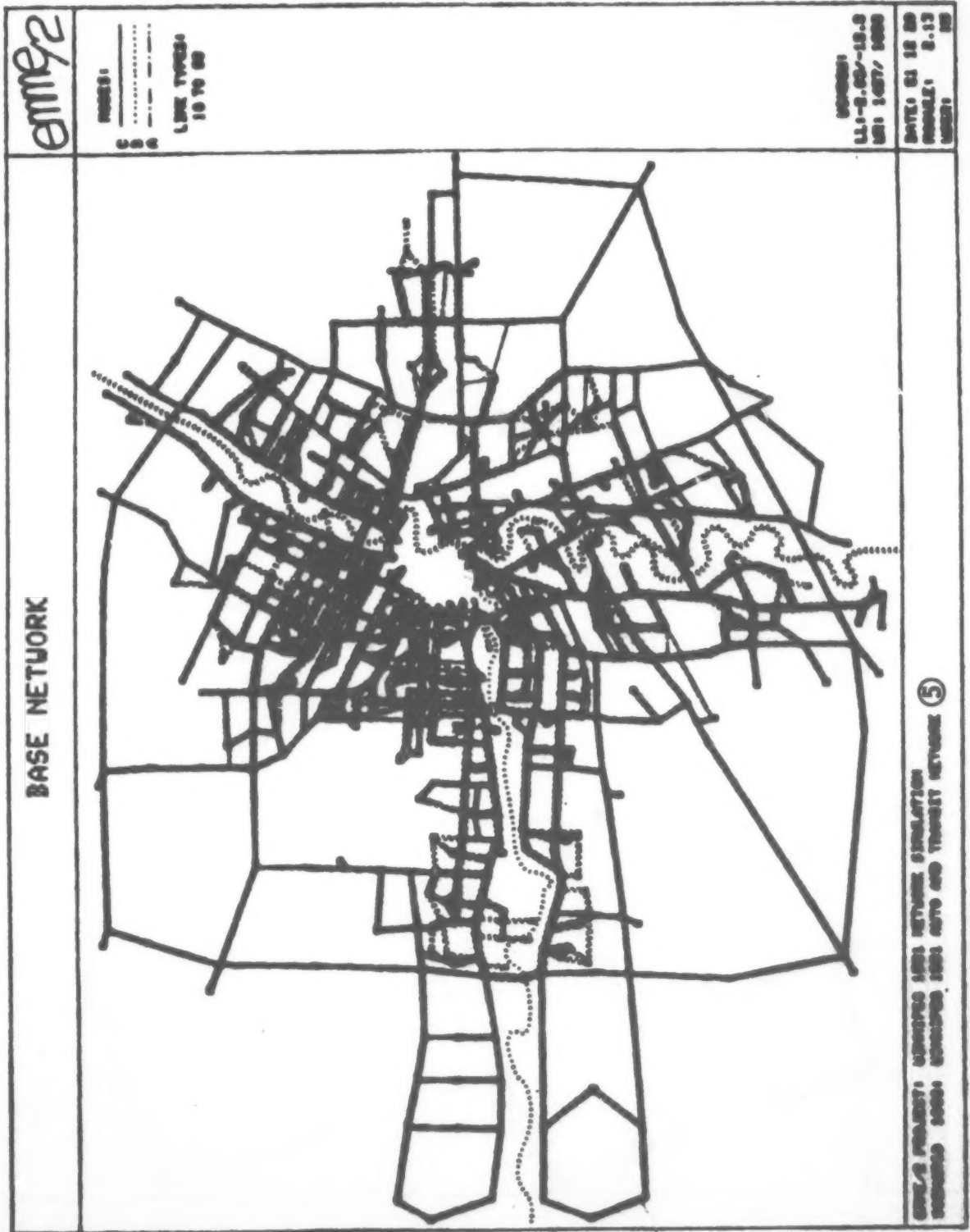
5. ASSIGNMENT PROCEDURES

- 5.11 PREPARE SCENARIO FOR ASSIGNMENT
- 5.21 AUTO ASSIGNMENT
- 5.31 TRANSIT ASSIGNMENT
- 5.41 BIMODAL ASSIGNMENT

6. RESULTS

- 6.01 RESULT SUMMARY
- 6.11 LIST AUTO TIMES AND VOLUMES
- 6.12 PLOT AUTO TIMES AND VOLUMES
- 6.13 COMPARE AUTO TIMES AND VOLUMES
- 6.14 PLOT AUTO TIMES AND VOLUMES ON INTERSECTIONS
- 6.15 PLOT SHORTEST PATHS ON AUTO NETWORK
- 6.21 LIST TRANSIT VOLUMES
- 6.22 PLOT TRANSIT VOLUMES
- 6.23 COMPARE TRANSIT VOLUMES
- 6.31 LIST MODAL SPLIT
- 6.32 PLOT MODAL SPLIT
- 6.33 COMPARE MODAL SPLIT

Figure 2



⑤

Figure 3

2.13 PLOT BASE NETWORK

SELECT: TYPE OF DISPLAY

1= BASE NETWORK

2= MODES

3= LINK TYPES

4= LINK LENGTH

5= LANES / VOLUME-DELAY FUNCTIONS

6= USER DEFINED LINK DATA

7= MODIFY MODULE PARAMETERS

8= END

S

DISPLAY MODE NUMBERS? yes

ENTER: DESIRED LINK TYPES (FROM, TO)

=1,6

=/

WINDOW? yes

SELECT: 1= USING COORDINATES

2= USING MODES

3= CENTERED ON A MODE

4= USE SAME WINDOW AS BEFORE

3

ENTER: CENTER MODE (C-CENTER OF DEFAULT WINDOW)=947

ENTER: DIAMETER OF WINDOW (192.2)=25

Figure 4

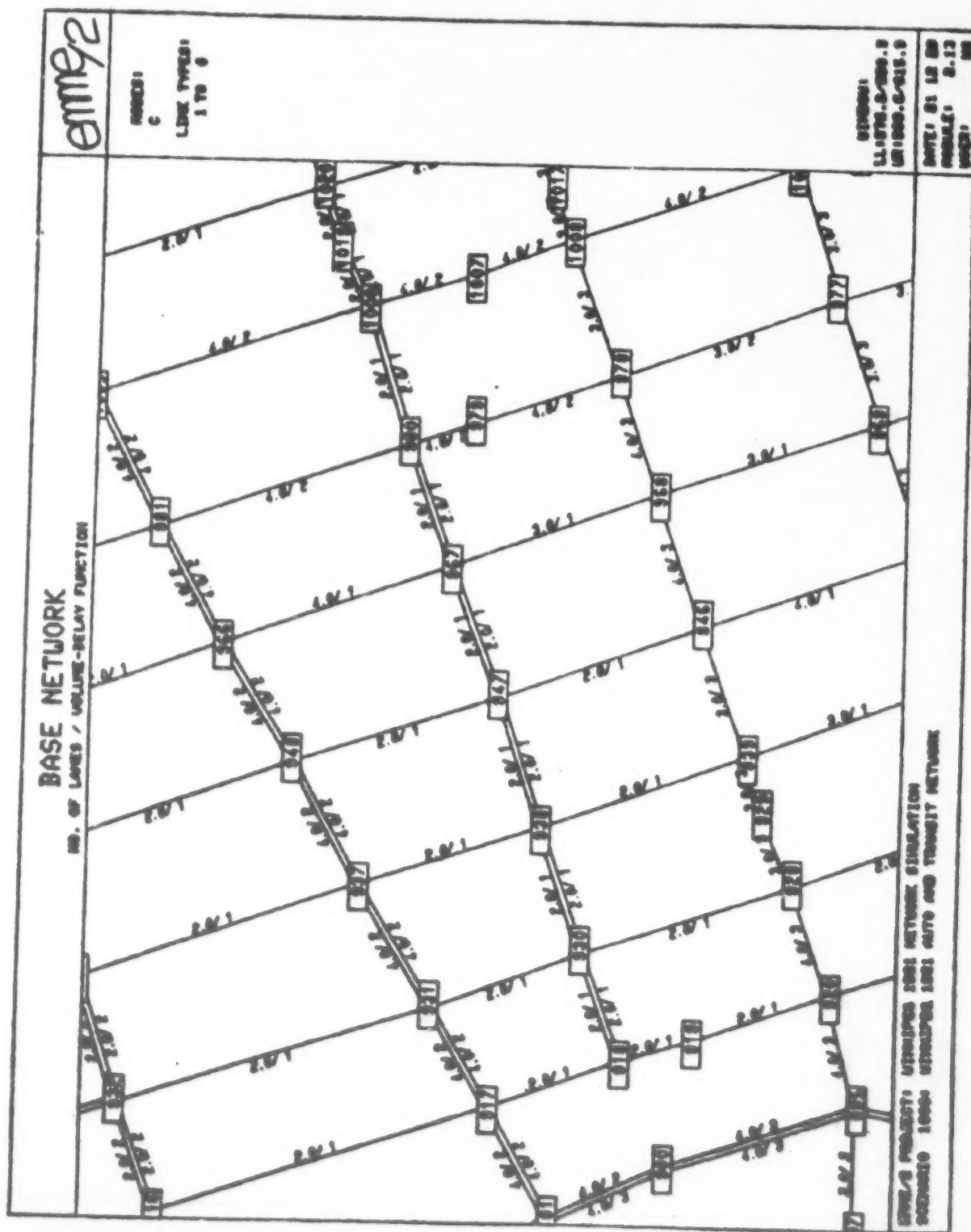


Figure 5

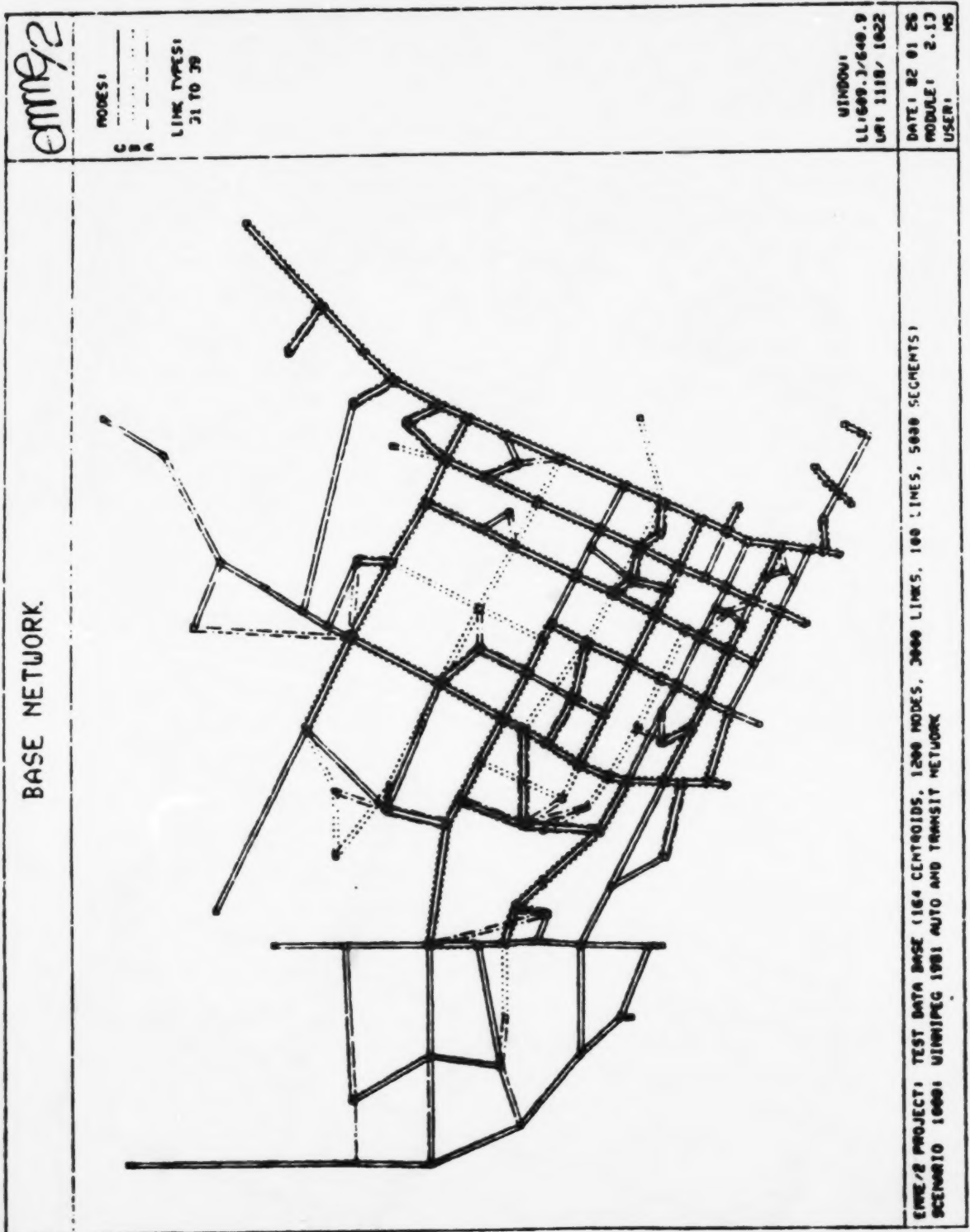


Figure 6

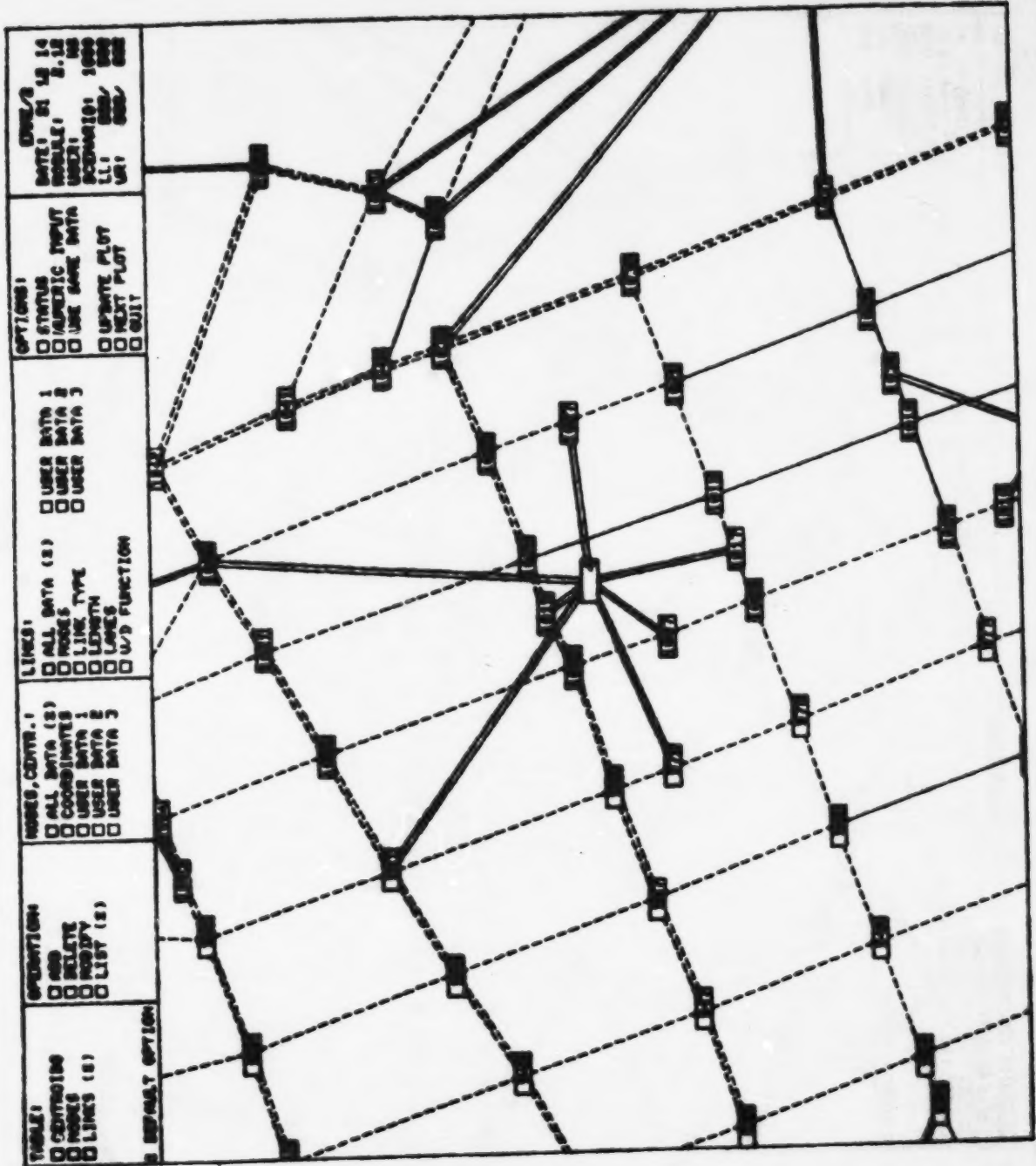
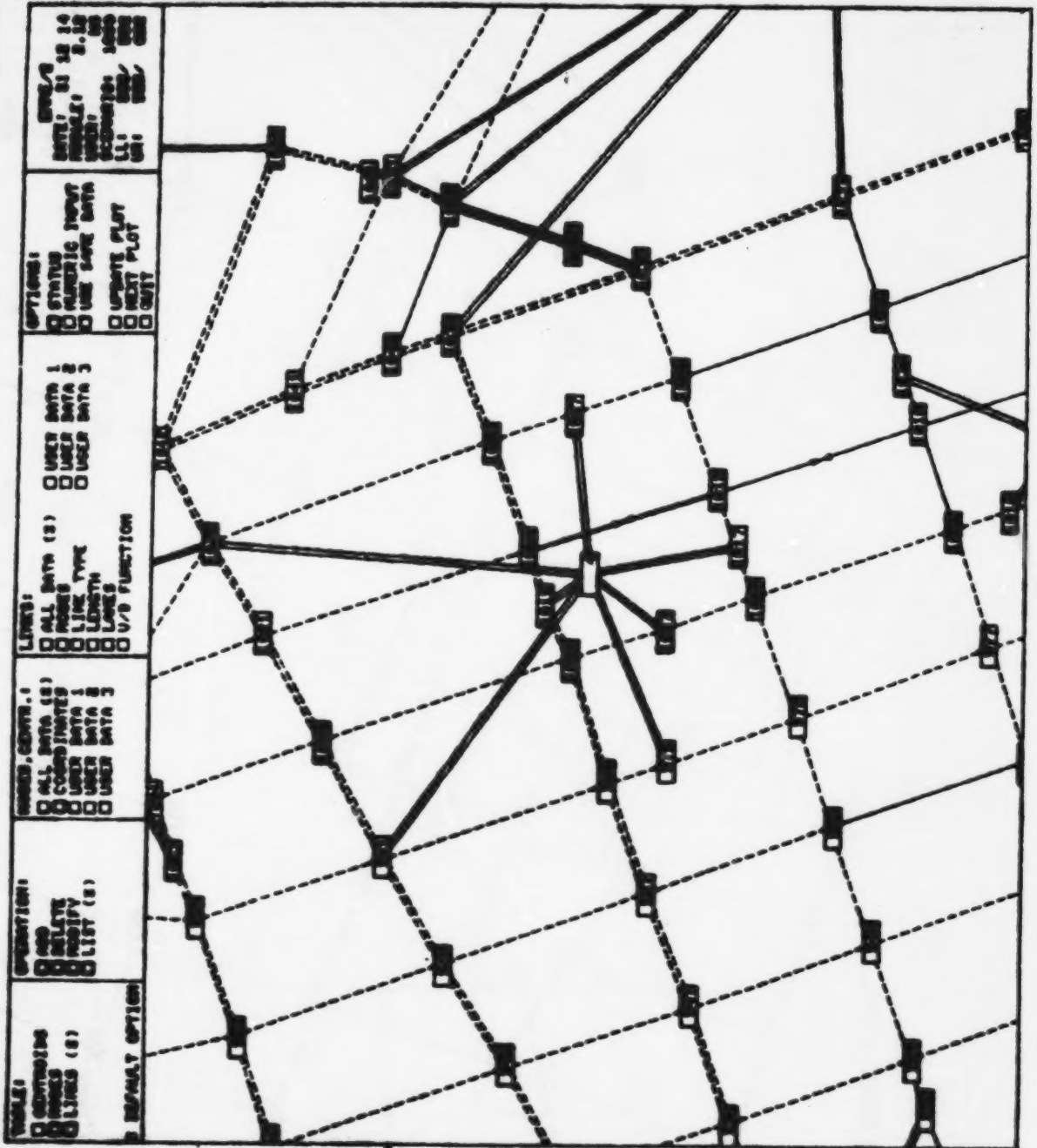


Figure 7



McGraw-Hill

Figure 8

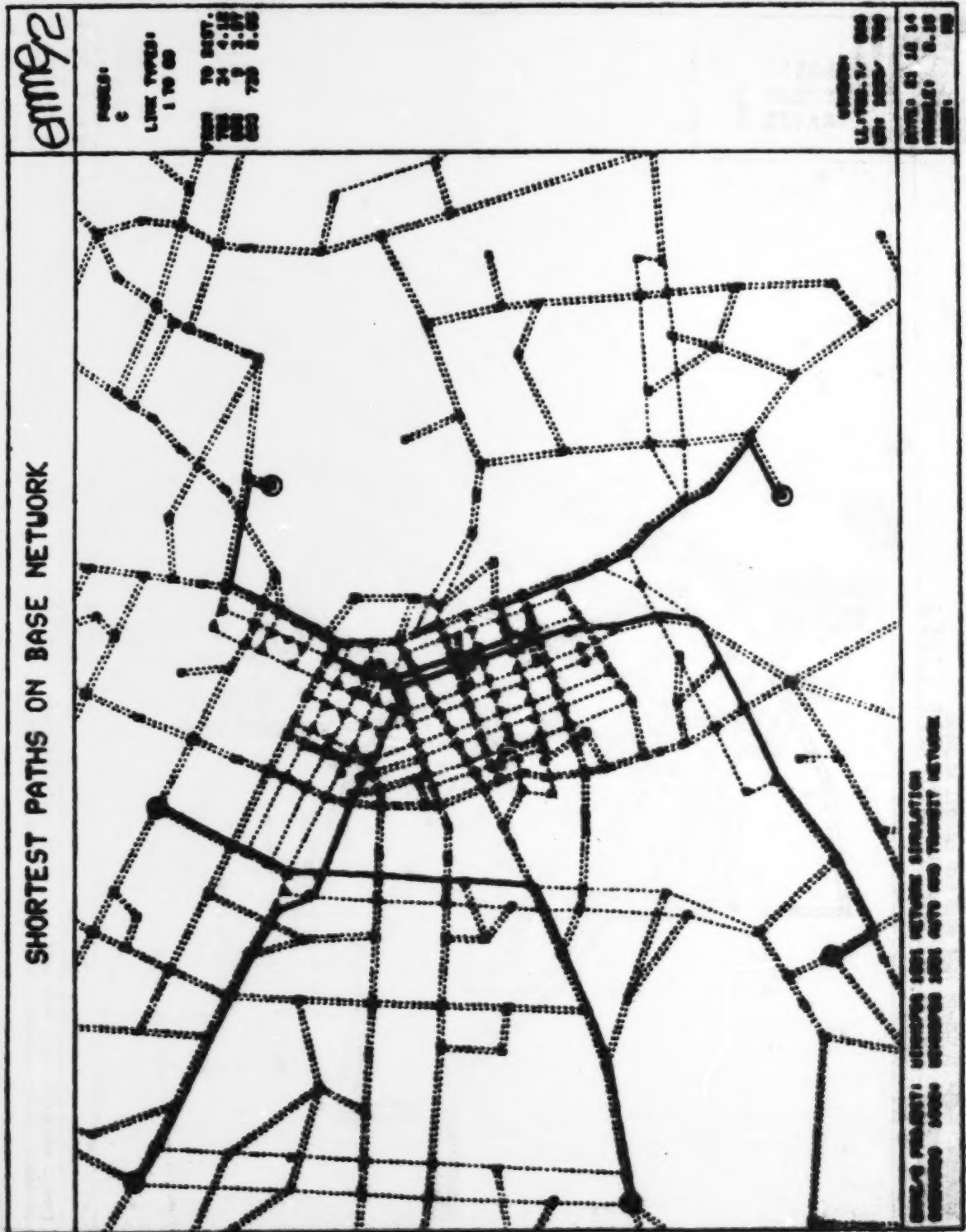


Figure 9

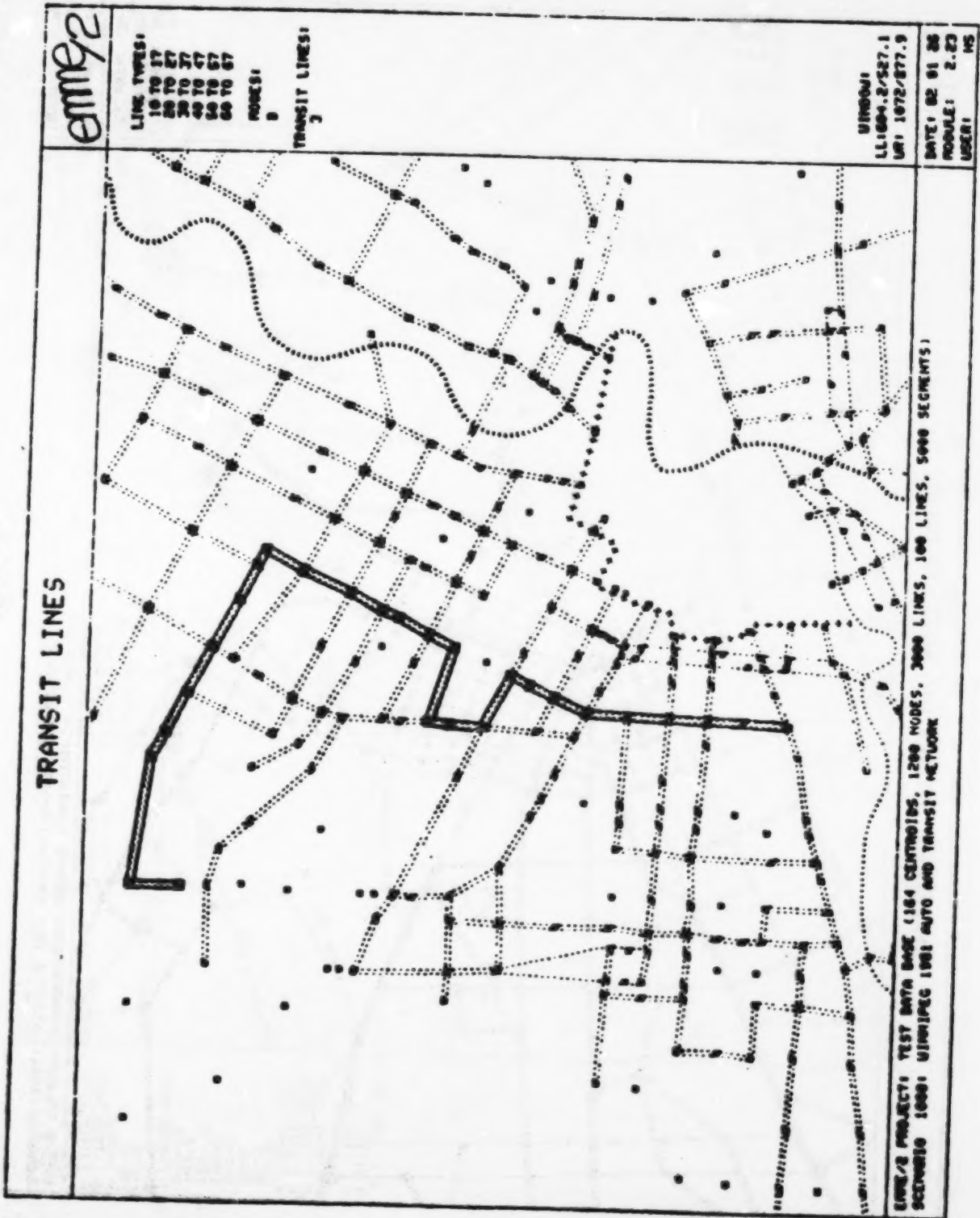


Figure 10

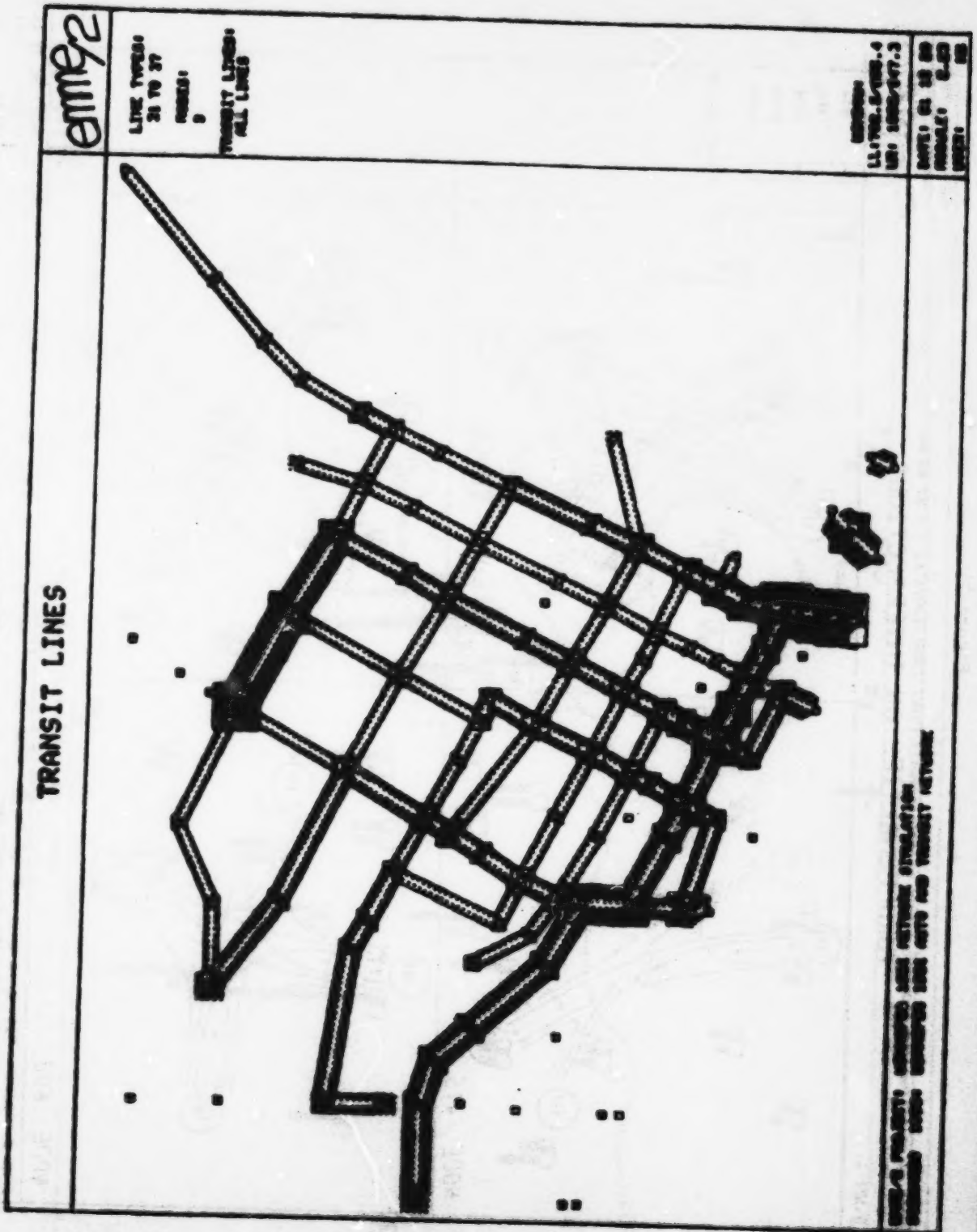


Figure 13

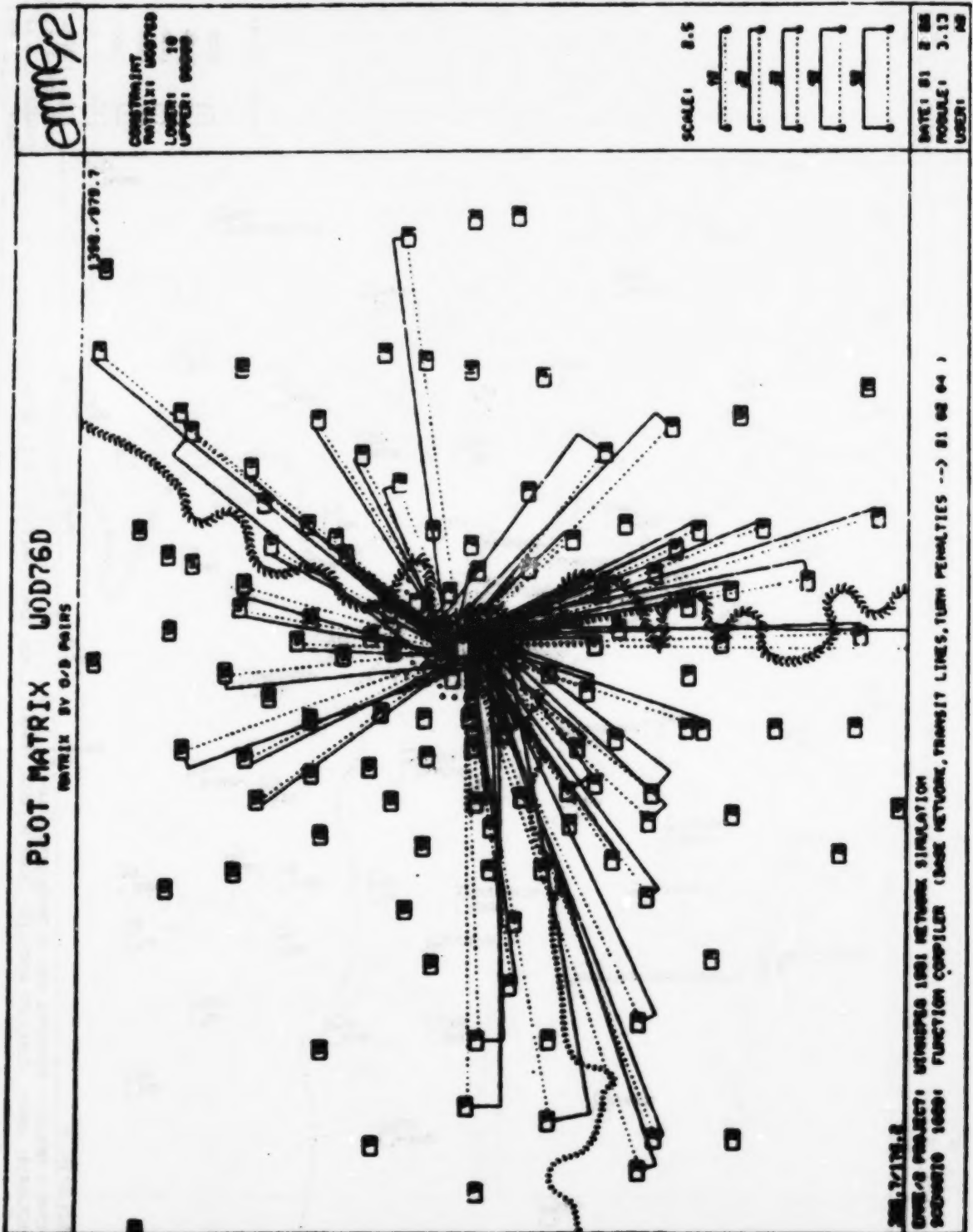


Figure 14

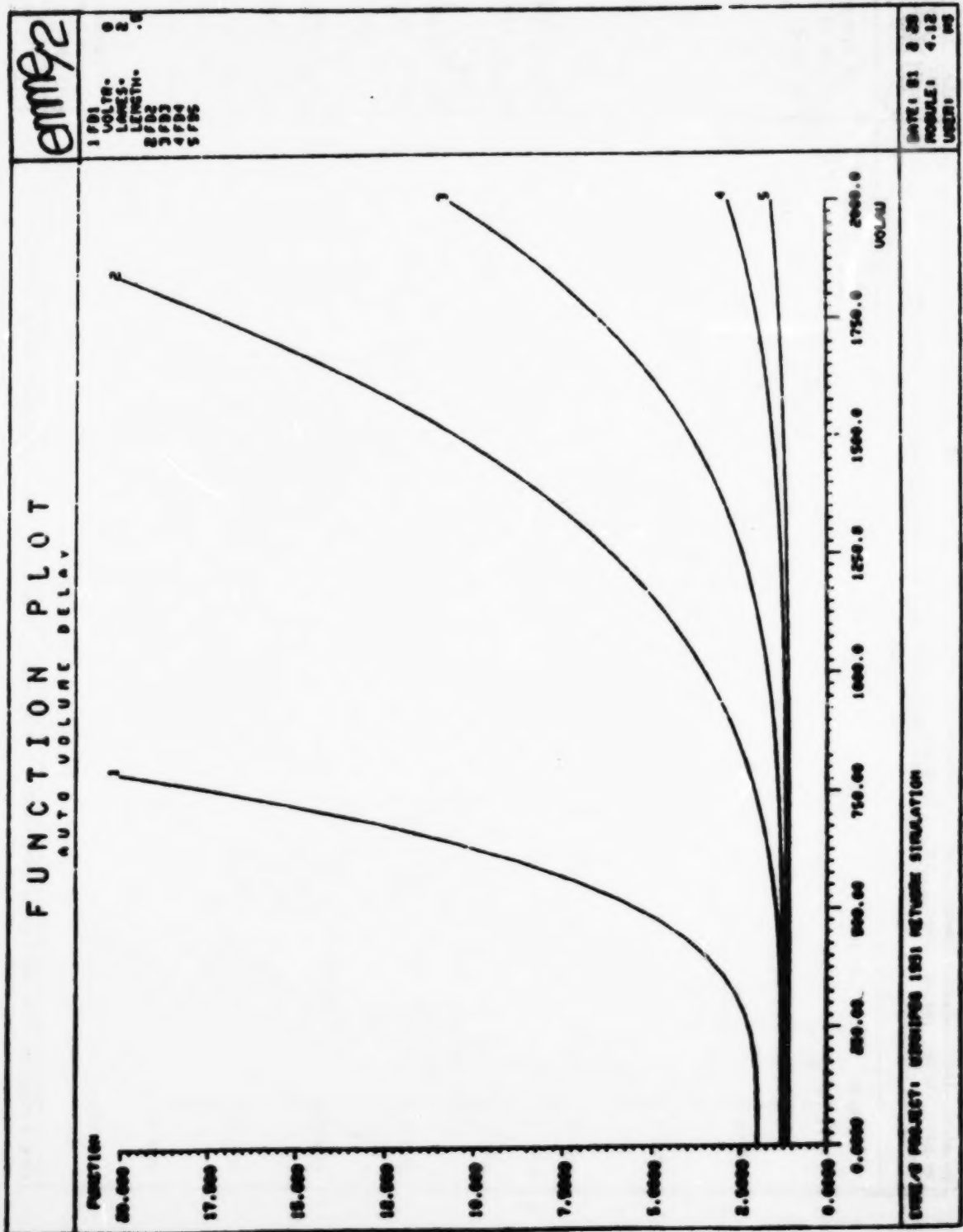


Figure 15

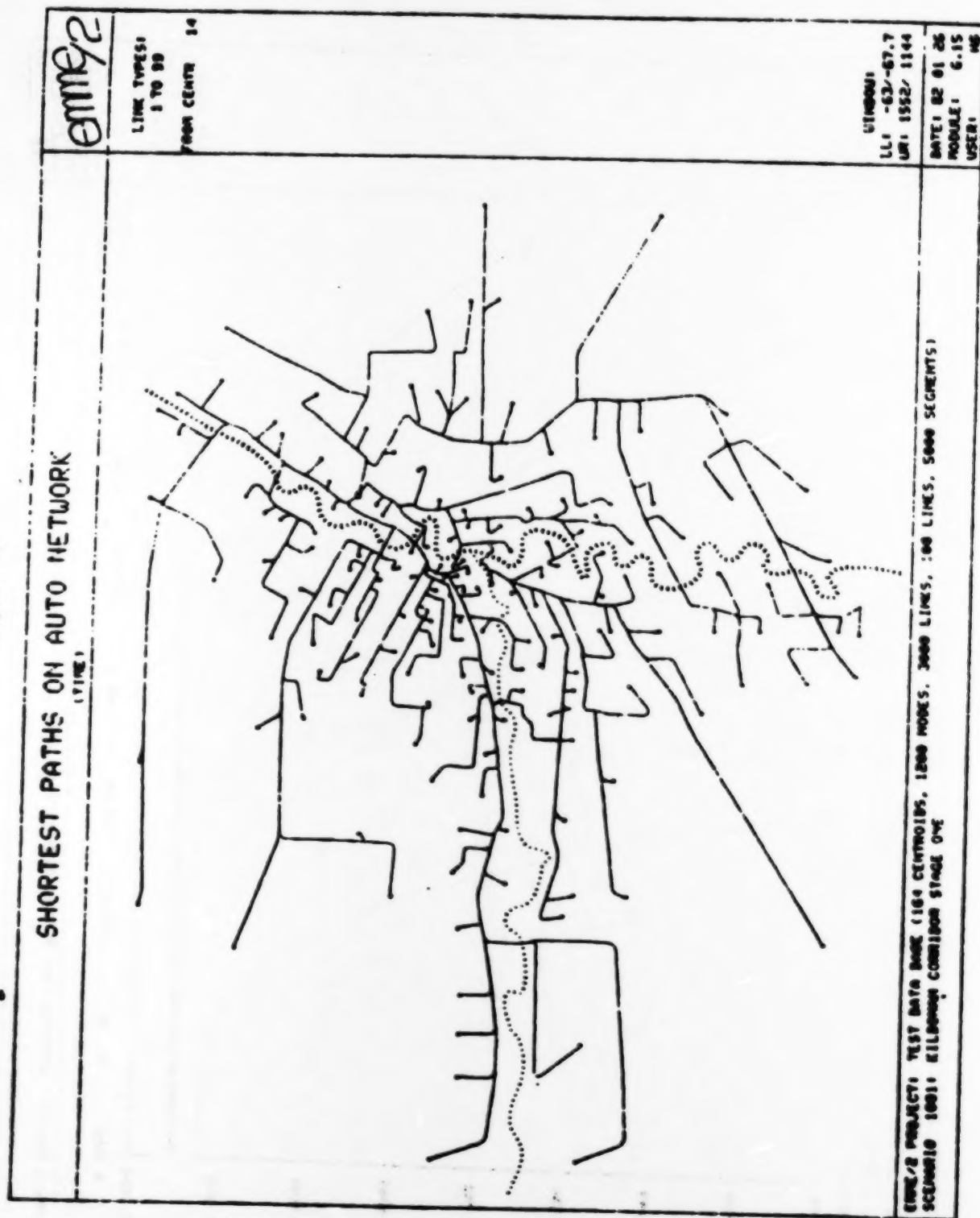


Figure 16

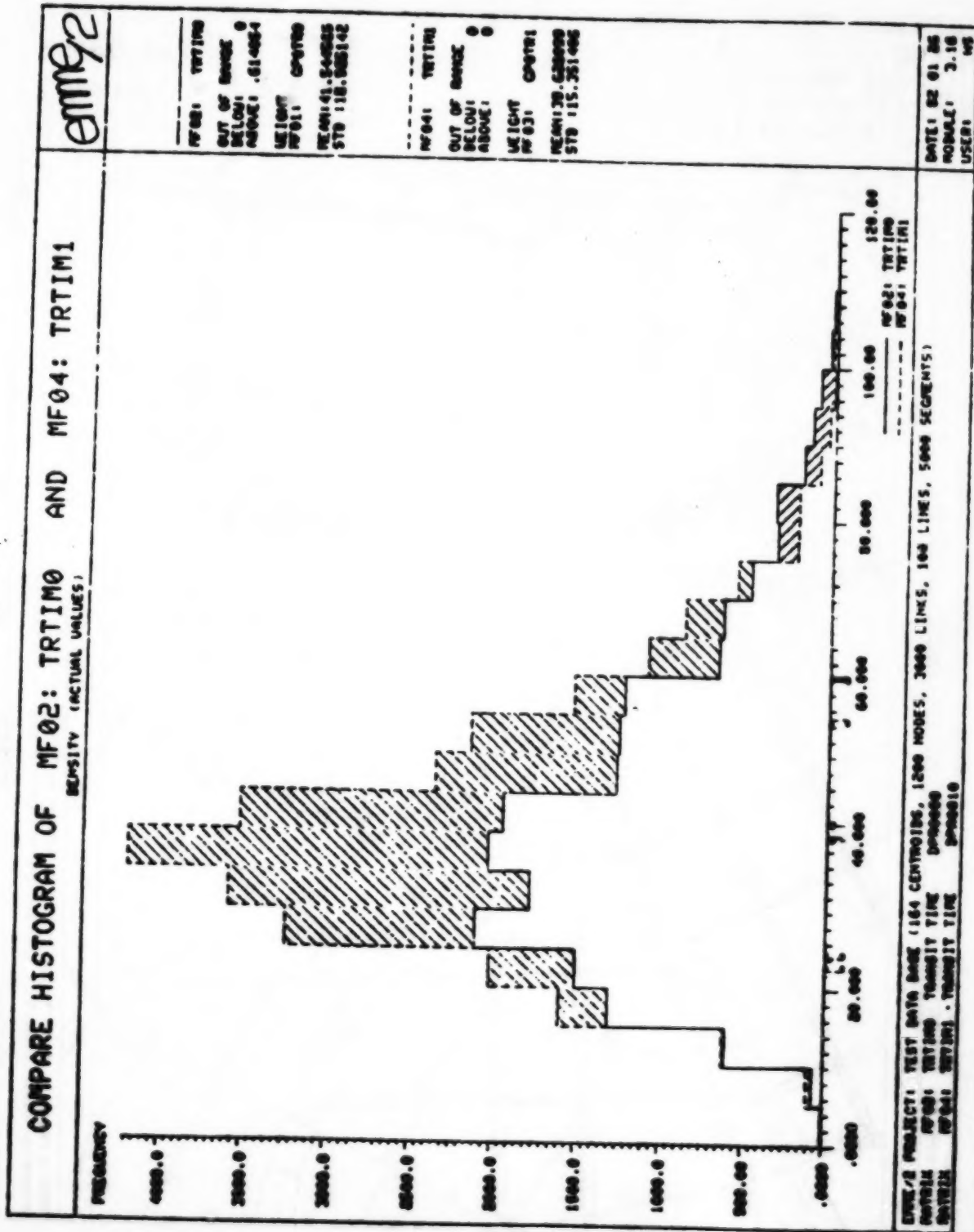


Figure 17

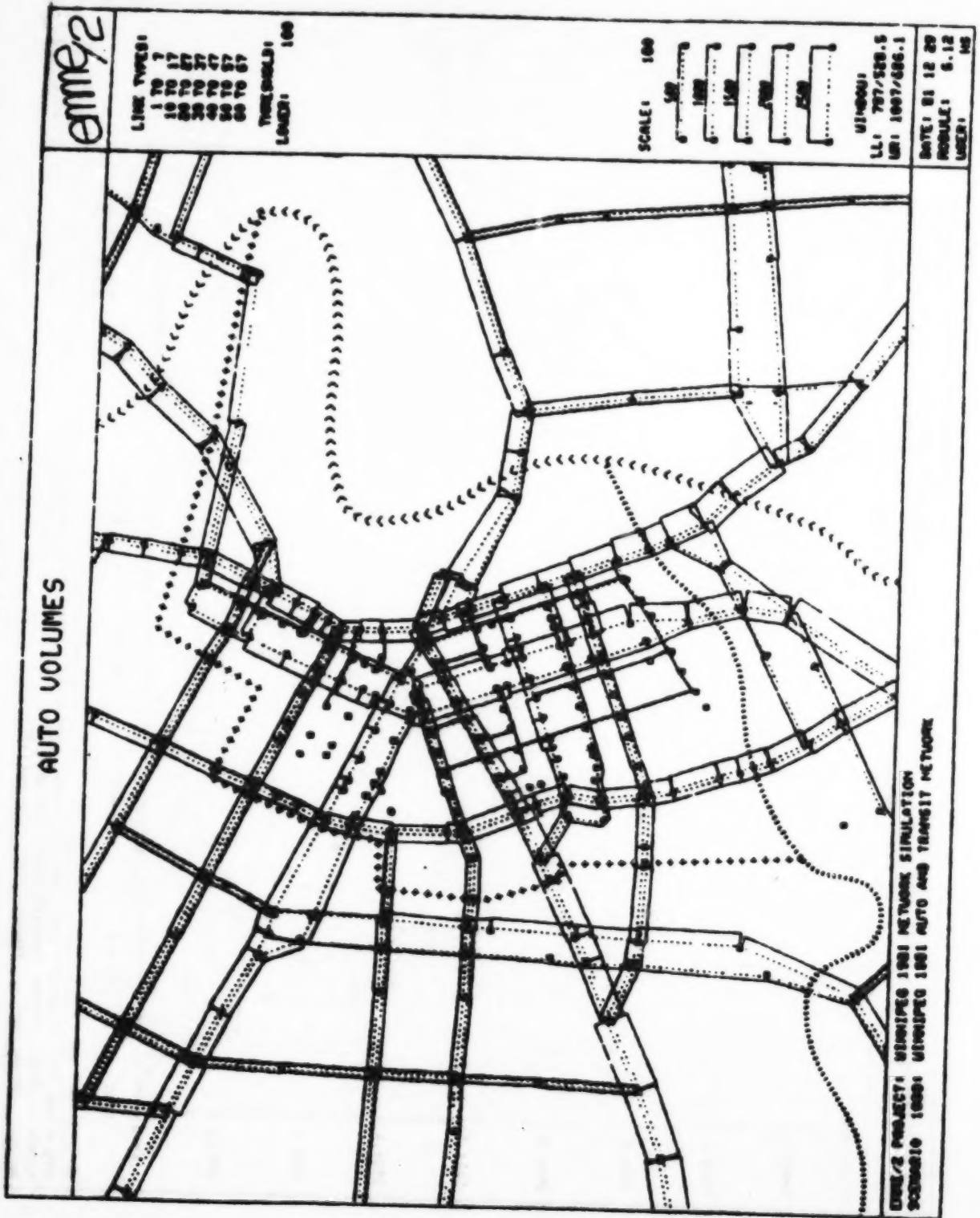
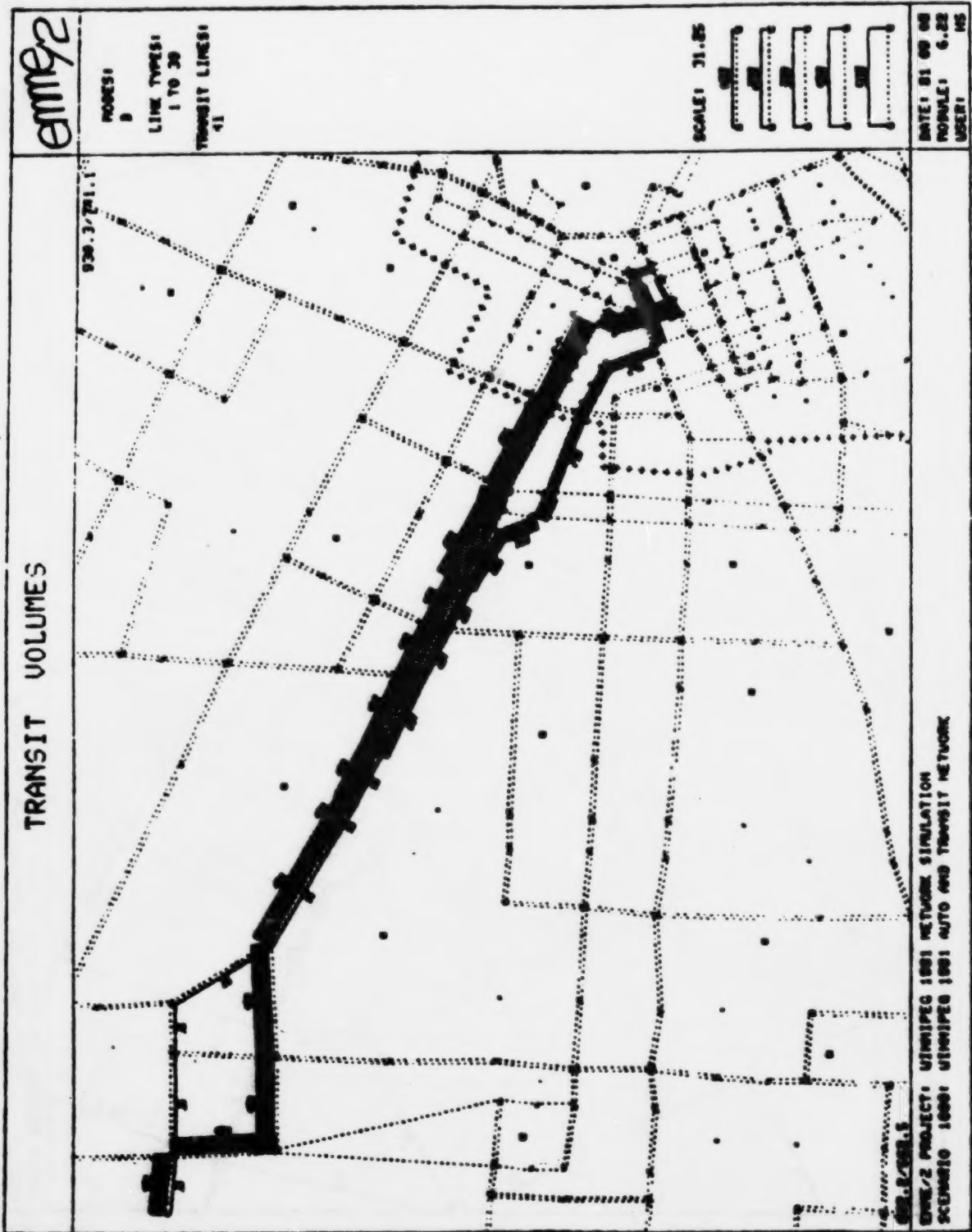


Figure 19



REFERENCES

- [1] H.Z. Aashtiani and T.L. Magnanti, "Equilibria on a Congested Transportation Network." Opns. Res. Ctr., To appear in SIAM Journal on Algebraic and Discrete Methods. M.I.T., (1980).
- [2] M.J. Beckmann, C.B. McGuire, and C.B. Winsten, Studies in the Economics of Transportation. Yale University Press, New Haven, CT, (1956).
- [3] D. Braess, Über ein Paradoxon der Verkehrsplanung. Unternehmensforschung, (1968), 258-268.
- [4] M. Bruynooghe, A. Gibert, and M. Sakarovitch, "Une méthode d'affectation du trafic," In Fourth Symposium on the Theory of Traffic Flow, Karlsruhe, (1968).
- [5] M.F. Chang, L. Evans, R. Herman, and P. Wasiliewski, "Gasoline Consumption in Urban Traffic," Transportation Research Record 599 (1976), 25-30.
- [6] S.C. Dafermos and A. Nagurney, Sensitivity Analysis for the General Asymmetric Network Equilibrium Problem. Technical Report LCDS # M-82-7, [Lefschetz Center for Dynamical Systems, Brown University, (1982)].
- [7] S.C. Dafermos, "Traffic Equilibrium and Variational Inequalities." Transportation Science (Vol. 14, 1980), 42-54.
- [8] S.C. Dafermos, "The General Multimodal Network Equilibrium Problem with Elastic Demand," Lefschetz Center for Dynamical Systems, Brown University, (1980). To appear in Networks.
- [9] S.C. Dafermos, "An Extended Traffic Assignment Model with Applications to Two-way Traffic," Transportation Science (4, 1971), 336-389.
- [10] G. Dantzig, S.F. Maier, and Z.F. Lansdowne, Application of Decomposition to Transportation Network Analysis. Tech. Report DOT-TSC-OST-76-26, Control Analysis Corporation, (1976).
- [11] R.B. Dial, "A Probabilistic Multipath Assignment Model which Obviates Path Enumeration," Transportation Research (8, 1971), 85-96.
- [12] P. Dow and D. Van Vliet, "Capacity Restrained Road Assignment," Traffic Engineering and Control (1979), 261-273.
- [13] J. Dupuit, "De la mesure de l'utilité des travaux publics," Annales des travaux publics, Annales des ponts et chaussées, (2e série, 8, 1844).
- [14] S.P. Evans, "Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment," Transportation Research (10, 1976), 37-57.

- [15] C. Fisk and S. Pallottino, "Empirical Evidence for Equilibrium Paradoxes with Implications for Optimal Planning Strategies," Transportation Research (Vol. 15A, No. 3, 1981), 245-248.
- [16] M. Florian and S. Nguyen, "An Application and Validation of Equilibrium Trip Assignment Methods," Transportation Science (Vol. 10, 1976), 374-389.
- [17] M. Florian, J. Ferland and S. Nguyen, "On the Combined Distribution-Assignment of Traffic," Transportation Science (9, 1975), 43-53.
- [18] M. Frank and P. Wolfe, "An Algorithm of Quadratic Programming," Naval Res. Log. Quart. (3, 1956), 95-110.
- [19] C. Frank, "A Study of Alternative Approaches to Combined Trip Distribution-Assignment Modeling," (Ph.D. Thesis, Department of Regional Science, University of Pennsylvania, 1978).
- [20] M. Hall, "Properties of the Equilibrium State in Transportation Networks," Transportation Science (Vol. 12, 1978), 208-216.
- [21] D. Kinderlehrer and G. Stampacchia, An Introduction to Variational Inequalities and Applications. (Academic Press, 1980).
- [22] F.H. Knight, "Some Fallacies in the Interpretation of Social Cost," Quarterly Journal of Economics (38, 1924), 582-606.
- [23] L.J. LeBlanc and K. Farhangian, "Efficient Algorithms for Solving Elastic Demand Traffic Assignment Problems and Mode Split Assignment Problems," Transportation Science (15, 1981), 306-312.
- [24] L.J. LeBlanc, E.K. Morlok and W.P. Pierskalla, "An Efficient Approach to Solving the Road Network Equilibrium Traffic Assignment Problem," Transportation Research (5, 1975), 309-318.
- [25] L.J. LeBlanc, "Mathematical Programming Algorithms for Large Scale Network Equilibrium and Network Design Problems," (Ph.D. Dissertation, Dept. of Ind. Eng. and Mgt. Sci., Northwestern University, 1973).
- [26] J.D. Murchland, "Braess' Paradox of Traffic Flow," Transportation Research (4, 1970), 391-394.
- [27] M. Netter, "Equilibrium and Marginal Cost Pricing on a Road Network with Several Traffic Types," In Traffic Flow and Transportation. Proceedings of the Fifth International Symposium on the Theory and Traffic Flow and Transportation, G.F. Newell (ed.), Springer-Verlag, (1971).
- [28] S. Nguyen, Procedures for Equilibrium Traffic Assignment with Elastic Demand. Publication #39, Centre de recherche sur les transports, Université de Montréal, (1977).

- [29] S. Nguyen, "A Unified Approach to Equilibrium Methods for Traffic Assignment," In Traffic Equilibrium Methods. Proceedings (1974) [M. Florian, Editor], Vol. 118 Lecture Notes in Economics and Mathematical Systems, (Springer Verlag, 1976), 148-182.
- [30] S. Nguyen, "Une approche unifiée des méthodes d'équilibre pour l'affectation du trafic," [Ph.D. Dissertation, Département d'informatique et de recherche opérationnelle, Université de Montréal, (1973)].
- [31] R.B. Potts and R.M. Oliver, Flows on Transportation Networks. Academic Press, New York, (1972).
- [32] M.J. Smith, "Existence, Uniqueness and Stability of Traffic Equilibria," Transportation Research B (Vol. 13B, 1979), 295-304.
- [33] N.F. Stewart, Notes on the Mathematical Structure of Equilibrium Models. Technical Report LITH-MATH-12-79-8, Department of Mathematics, Linköping Institute of Technology, (1979).
- [34] J.G. Wardrop, Some Theoretical Aspects of Road Traffic Research, Proc. Inst. Civil Engineers, Part II, (1952), 325-378.



BIBLIOGRAPHY

Multi-Mode Models - Non Interactive - Non Graphic:

Florian, M.; Chapleau, R.; Nguyen, S.; Achim, C.; James-Lefebvre, L.; Galarneau, S.; Lefebvre, J.; and Fisk, C. "Validation and Application of an Equilibrium-Based Two-Mode Urban Transportation Planning Method (EMME)." Transportation Research Record, (No. 728, 1979), 14-22.

Urban Mass Transportation Administration. Urban Transportation Planning System Reference Manual, U.S. Department of Transportation, Washington, D.C., (1977) [and other updates].

Transit Planning - Interactive Graphic Applications:

Andreasson, I. "The Volvo Approach to Computer-Aided Transit Planning." Transportation Research Record, (No. 657, 1978).

General Motors Transportation Systems Center. Interactive Graphic Transit Design (IGTDS) Demonstration Study, Bellevue, Washington. Final Report, Volumes I and II (EP-80023A) GMTSC: Warren Michigan, (1980).

Kornhauser, A.L., and Hess, J.L. "Interactive Graphics Sketch Planning Model for Urban Transportation." Transportation Research Record (No. 647, 1978).

Rapp, M.H.; Piguet, S.; and Robert-Grandpierre, A. "Interactive Graphic System for Transit Route Optimization." Transportation Research Record (No. 588, 1978).

Riepl, R.J., and Gehner, C.D. "Interactive Graphics in Europe: Three Computer-Aided Transportation Tools." Computer Graphics (Vol. 2, No. 5, 1979).

Schneider, J.B. "Interactive Graphics in Transportation System Planning." DOT-TST-74-10, Final report on a seminar held in Seattle, Washington, (January 1974), 99, MTIS, Springfield, VA.

Assignment Algorithms

Abdulaal, M., and LeBlanc, L.J. "Methods for Combining Modal Split and Equilibrium Assignment Models." Transportation Science (Vol. 13, 1979), 292-314.

Chriqui, C., and Robillard, P. "Common Bus Lines." Transportation Science (Vol. 9, 1975), 115-121.

Dafermos, S.C. Relaxation Algorithms for the General Asymmetric Traffic Equilibrium Problem, Lefschetz Centre for Dynamical Systems, Brown University, (1981).

Fisk, C. and Nguyen, S. "Properties of the Two Mode Equilibrium Assignment Model with Flow Dependent Transit Costs." Transportation Science (Vol. 15, No. 4, 1981), 318-328.

Florian, M. "A Traffic Equilibrium Model of Travel by Car and Public Transit Modes." Transportation Science (Vol. 8, 1977), 169-179.

Florian, M., and Spiess, H. "On Two Mode Choice/Assignment Models." Publication #177, Centre de recherche sur les transports, Université de Montréal, (1981) [To appear in Transportation Science].

Florian, M., and Spiess, H. "The Convergence of Diagonalization Methods for Fixed Demand Asymmetric Network Equilibrium Problems." Publication #198, Centre de recherche sur les transports, Université de Montréal, (1981) [To appear in Transportation Research].

Matrix Balancing Algorithms

Evans, S.P., and Kirby, H.R. "A Three-Dimensional Furness Procedure for Calibrating Gravity Models." Transportation Research (Vol. 8, 1974).

Evans, E.W. "The Calibration of Trip Distribution Models with Exponential or Similar Cost Functions." Transportation Research (Vol. 5, 1971).

APPENDICES



Appendix I

CONVEX AND MONOTONE FUNCTIONS

DEFINITION: A numerical function f defined over a convex set Ω is convex if, for every $x, y \in \Omega$ and for $\alpha, 0 \leq \alpha \leq 1$

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha)y)$$

If the above inequality is satisfied as a strict inequality for $x \neq y, \alpha, 0 < \alpha < 1$, then the function f is strictly convex.

DEFINITION: A numerical function g defined on a convex set Ω is said to be concave if the function $f = -g$ is convex. The function g is strictly concave if $-g$ is strictly convex.

PROPERTY 1: Let f_1 and f_2 be convex functions defined over a convex set Ω . Then the function $f_1 + f_2$ is convex on Ω .

PROPERTY 2: Let f be a convex function over the convex set Ω .

The function αf is convex for any $\alpha \geq 0$.

There are several important and useful properties of differentiable convex functions.

PROPERTY 3: Let f be a differentiable convex function defined on a convex set Ω . f is convex if and only if $f(y) \geq f(x) + \nabla f(x)(y-x)$ for all $x, y \in \Omega$.

DEFINITION: A vector function h defined on a convex set Ω is monotone if $(h(x)-h(y))(x-y) \geq 0$ for $x, y \in \Omega$. If the above inequality is satisfied as a strict inequality for $x \neq y, x, y \in \Omega$, then the function is strictly monotone. If a function h is monotone its negative is antitone.

PROPERTY 4: Let f be a differentiable convex function defined on a convex set Ω . f is convex if and only if ∇f is monotone, that is $(\nabla f(x) - \nabla f(y))(x - y) \geq 0$ for $x, y \in \Omega$.

The following are classical results concerning the minimization of convex functions.

PROPERTY 5: Let f be a convex function defined on a convex set Ω . Then f achieves its minimum over a convex set Γ and any local minimum of f is a global minimum.

PROPERTY 6: Let f be a differentiable convex function defined on a convex set Ω . x^* is a global minimum of f over Ω if for all $y \in \Omega$, $\nabla f(x^*)(y - x^*) \geq 0$.

REFERENCES FOR APPENDIX I

O.L. Mangasarian. Nonlinear Programming. McGraw Hill, 1964.

D.G. Luenberger. Introduction to Linear and Nonlinear Programming. Addison-Wesley, 1965.

Appendix II

PRIMAL METHODS FOR NONLINEAR PROGRAMS

Consider the solution of the nonlinear programming problem

$$\min f(x)$$

$$\text{subject to } g(x) \leq 0$$

$$h(x) = 0$$

A primal method of solution is a procedure that works on the problem directly by searching through the feasible region for the optimal solution. Each point generated by the procedure is feasible and the value of the objective function decreases constantly.

A feasible direction d_k is a vector that permits, given a feasible x_k^l to generate $x^{l+1} = x^l + \lambda^l d^l$, where λ^l is a nonnegative scalar. This scalar is chosen to minimize the objective function f such that x_{k+1} is feasible. Thus a vector d_k is a feasible direction at x_k if there is a $\bar{\lambda} > 0$ such that $x_k + \lambda d_k$ is feasible for all λ , $0 \leq \lambda \leq \bar{\lambda}$ and $\nabla f(x^l)^T d^l \leq 0$.

One of the simplest feasible direction methods for the problem of minimizing a convex function f subject to the linear constraints $Ax = b$, $x \geq 0$, is the linear approximation method. Given a feasible point x_k , the direction vector is chosen to be $(y^l - x^l)$ where y^l is the solution of the linear program

$$\min \nabla f(x^l)^T y$$

$$\text{subject to } Ay = b$$

$$y \geq 0$$

$d^l = (y^l - x^l)$ is a direction of descent since the optimality conditions of the above linear program are that $\nabla f(x^l)^T (y - y^l) \geq 0$ and for $y = x^l$ it follows that $\nabla f(x^l)^T (y^l - x^l) < 0$ unless $x^l = x^*$, the solution of the convex program, where $\nabla f(x^*)^T (y - x^*) \geq 0$.

The detailed steps of the linear approximation algorithm for solving this problem are:

STEP 1- Given a feasible solution x^1 , set $l = 1$.

STEP 2- Determine y^l that minimizes

$$\nabla f(x^l)^T y, \text{ subject to } Ay = b, y \geq 0.$$

STEP 3- Set the descent direction $d^l = y^l - x^l$.

If $|\nabla f(x^l)^T d^l| \leq \epsilon$, terminate, where ϵ is a suitable convergence parameter, (x^l is the optimal solution).

STEP 4- Find the optimal step length λ^l that minimizes $F(x^l + \lambda d^l)$ for $0 \leq \lambda \leq 1$.

STEP 5- Revise the current solution $x^{l+1} = x^l + \lambda^l d^l$;

set $l = l + 1$ and return to Step 2.

Appendix III

THE KUHN-TUCKER CONDITIONS

Consider the nonlinear programming problem

$$\text{Min } f(x)$$

$$\text{subject to } g(x) \leq 0$$

$$h(x) = 0$$

Let x^* be a feasible solution which is a local minimum of the above problem. Then there exists a vector λ and a vector $\mu, \mu \geq 0$, such that

$$\nabla f(x^*) + \lambda \nabla h(x^*) + \mu \nabla g(x^*) = 0$$

and

$$\mu g(x^*) = 0$$

If $f(x)$ is convex and the feasible set is convex as well, the Kuhn-Tucker conditions are necessary and sufficient for x^* to be a global minimum of $f(x)$.

Appendix IV

DUALITY THEORY FOR LINEAR PROGRAMS

Consider the linear program

$$\begin{array}{ll}\text{Min } & cx \\ \text{subject to } & Ax = b \\ & x \geq 0\end{array}$$

and its corresponding dual problem

$$\begin{array}{ll}\text{Max } & ub \\ \text{subject to } & uA \leq c\end{array}$$

PROPERTY 1: If x and u are feasible primal and dual solutions and $cx = ub$, then they are optimal solutions as well.

PROPERTY 2: If x and u are feasible solutions for the primal and dual problems, a necessary and sufficient condition that they both be optimal is that $(uA - c)x = 0$.

REFERENCE FOR APPENDICES II, III AND IV

D.G. Luenberger. Introduction to Linear and Nonlinear Programming.
Addison-Wesley, 1965.